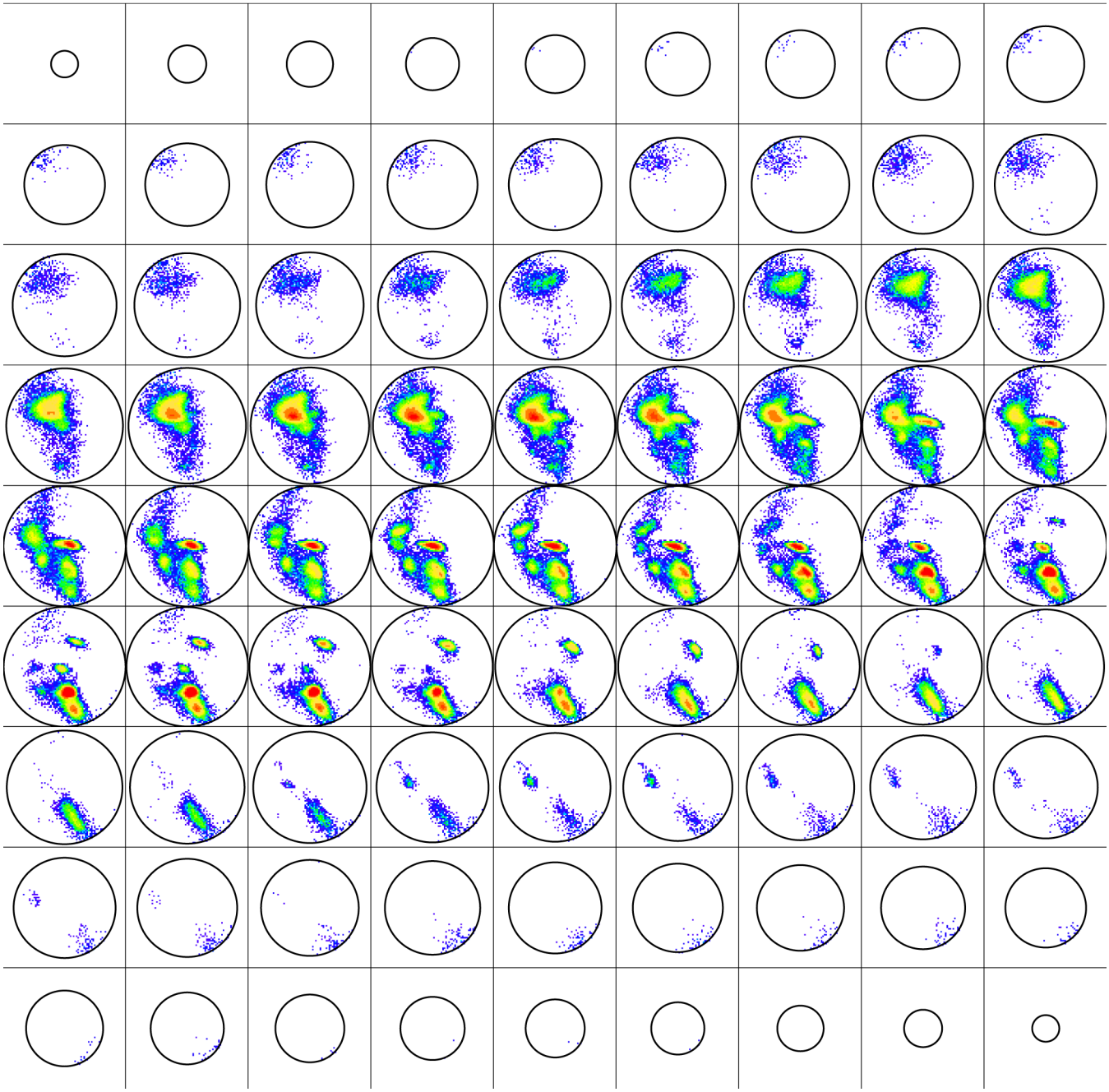
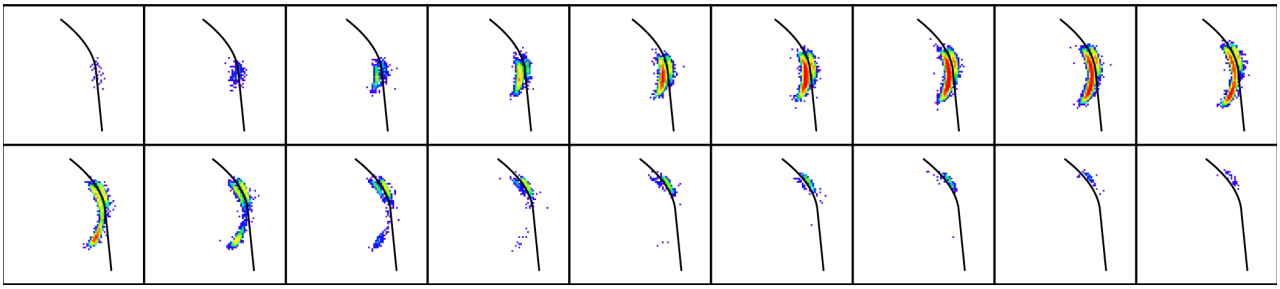
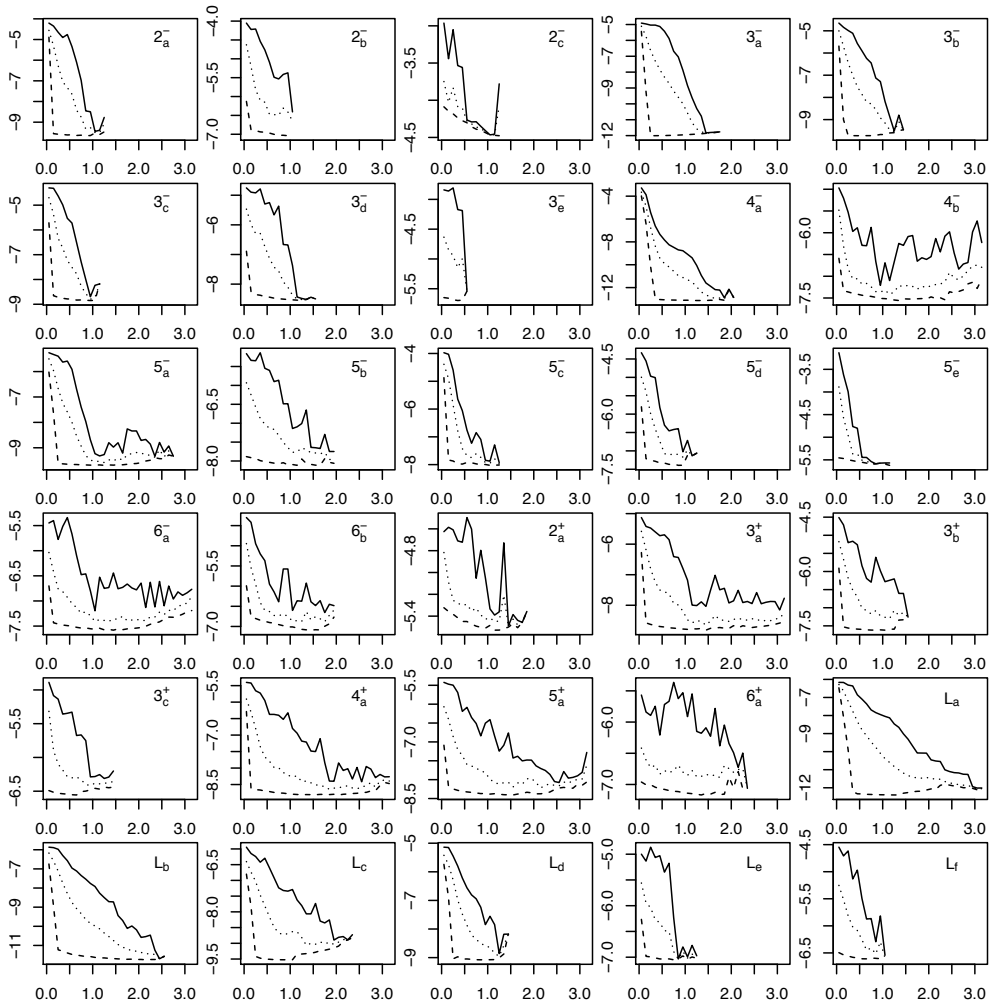
Supplementary Information



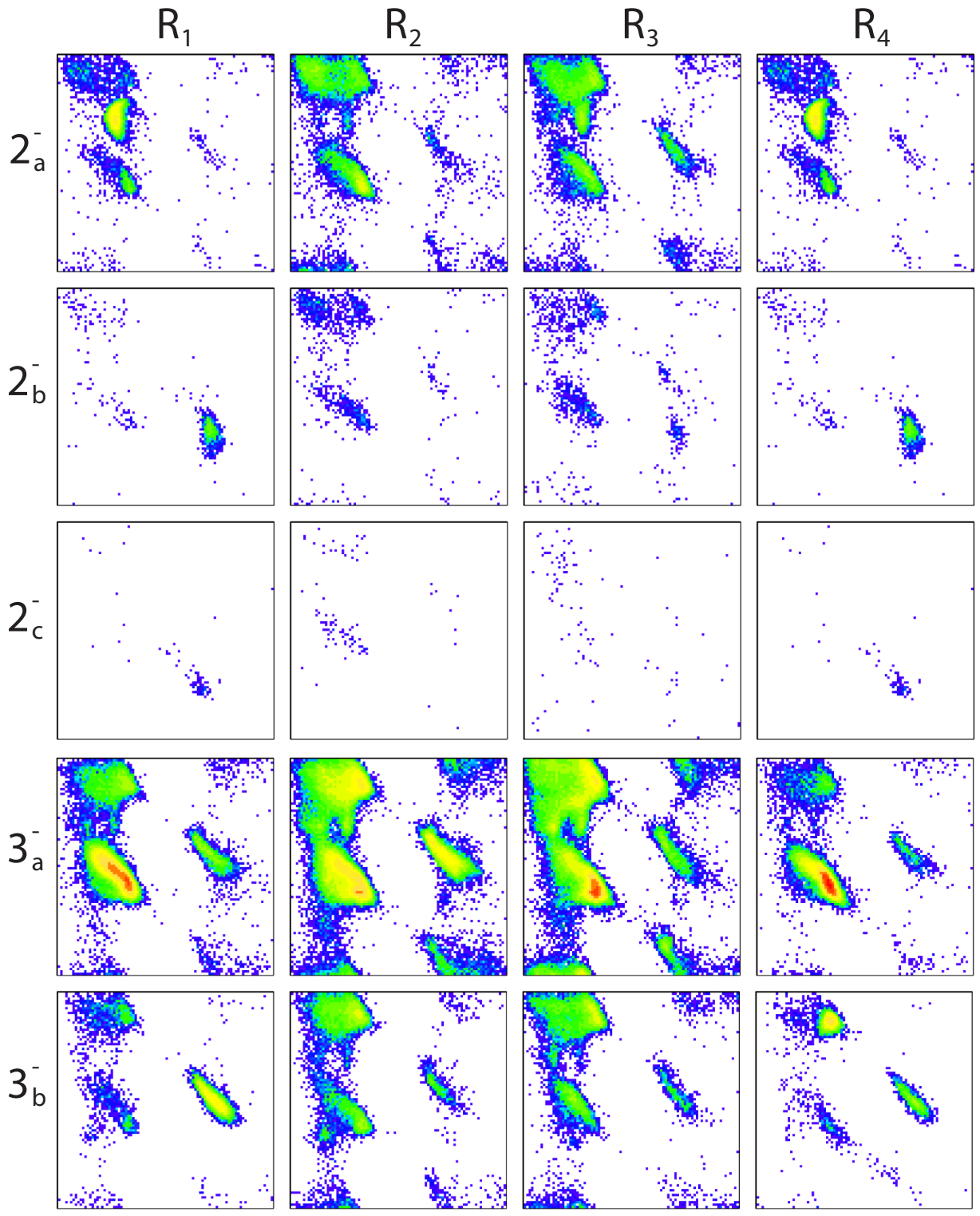
Supplementary Figure 1. Heat-plot histogram in rotational space of the HQ60 dataset. 81 horizontal slices from north to south pole in rotational space coloured by population density. The characteristic clustering of data is already visible from this heat-plot which should be compared to Figure 1 (B), (C), (D).



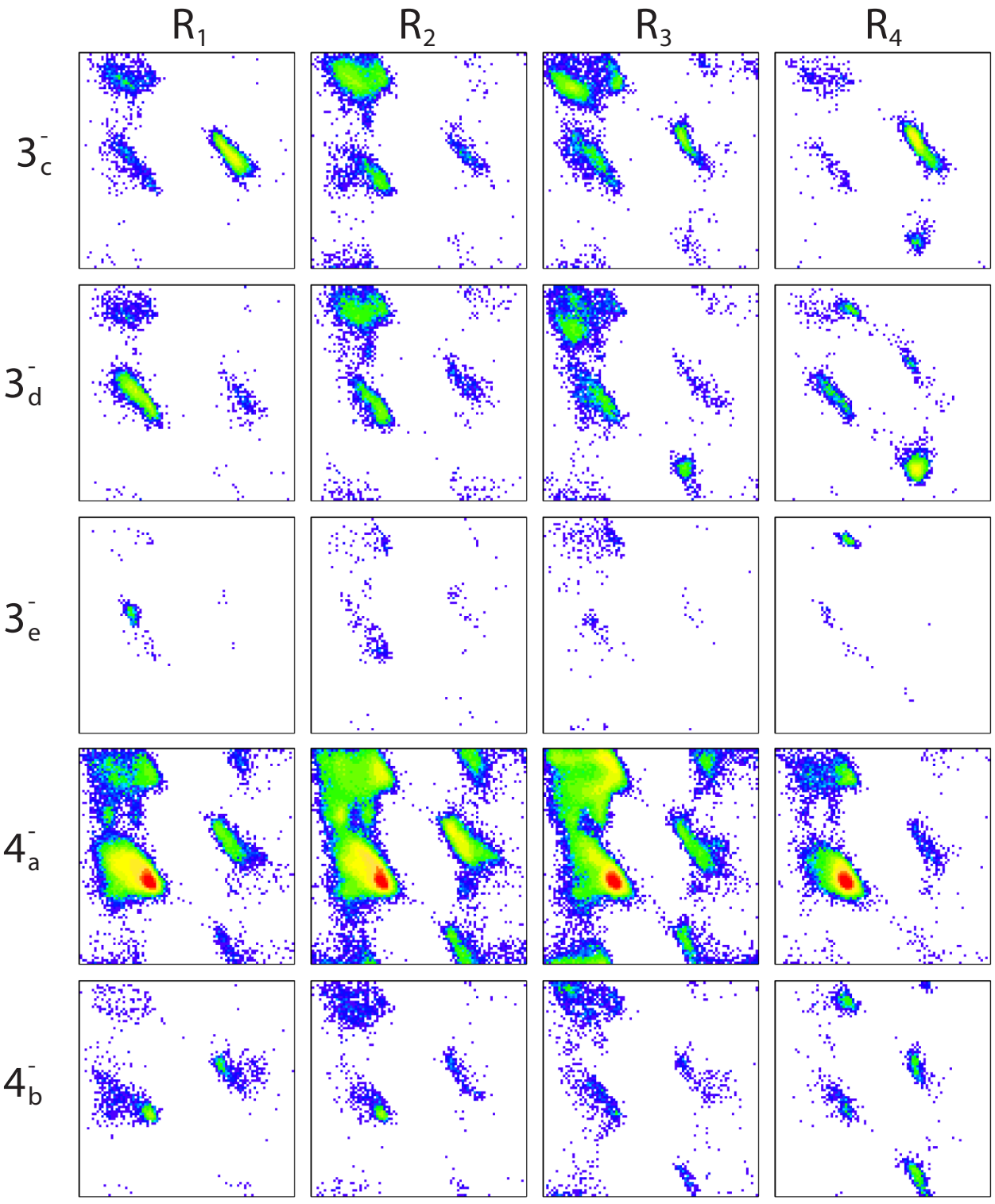
Supplementary Figure 2. Heat-plot of translations use to define 3_a^- and 3_b^-. Heat-plot of translations from the largest cluster of rotations of H-bonds with $\Delta = -3$. The translations concentrate in two clusters, which we separate with a surface, here illustrated in different slices with the black curves.



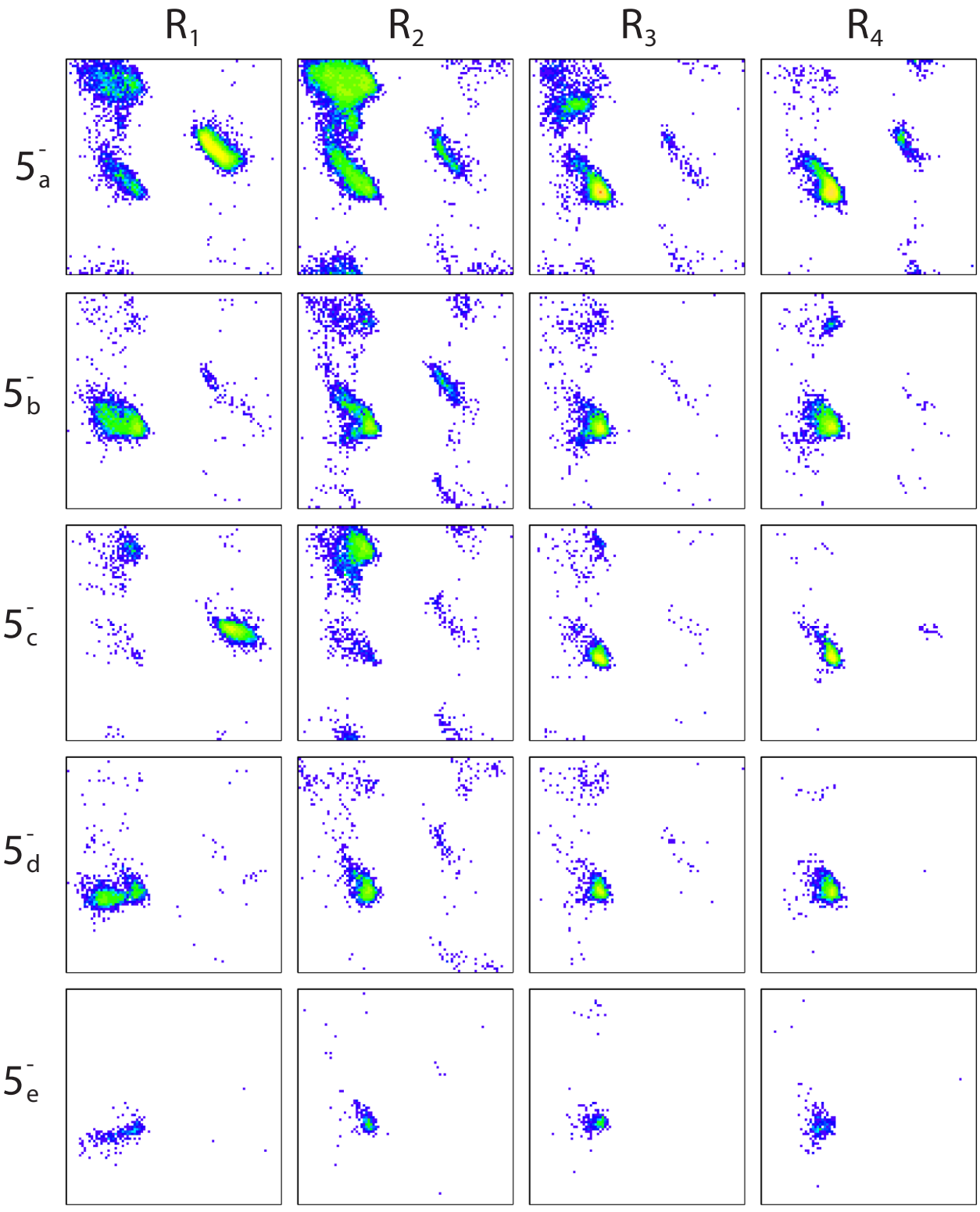
Supplementary Figure 3. Logarithmic density as a function of distance to mode. For each cluster the plot shows the maximal, minimal and average logarithmic density as a function of length to mode. Here the logarithmic density is normalised such that the density integrates to 1 within each cluster with respect to the Haar measure.



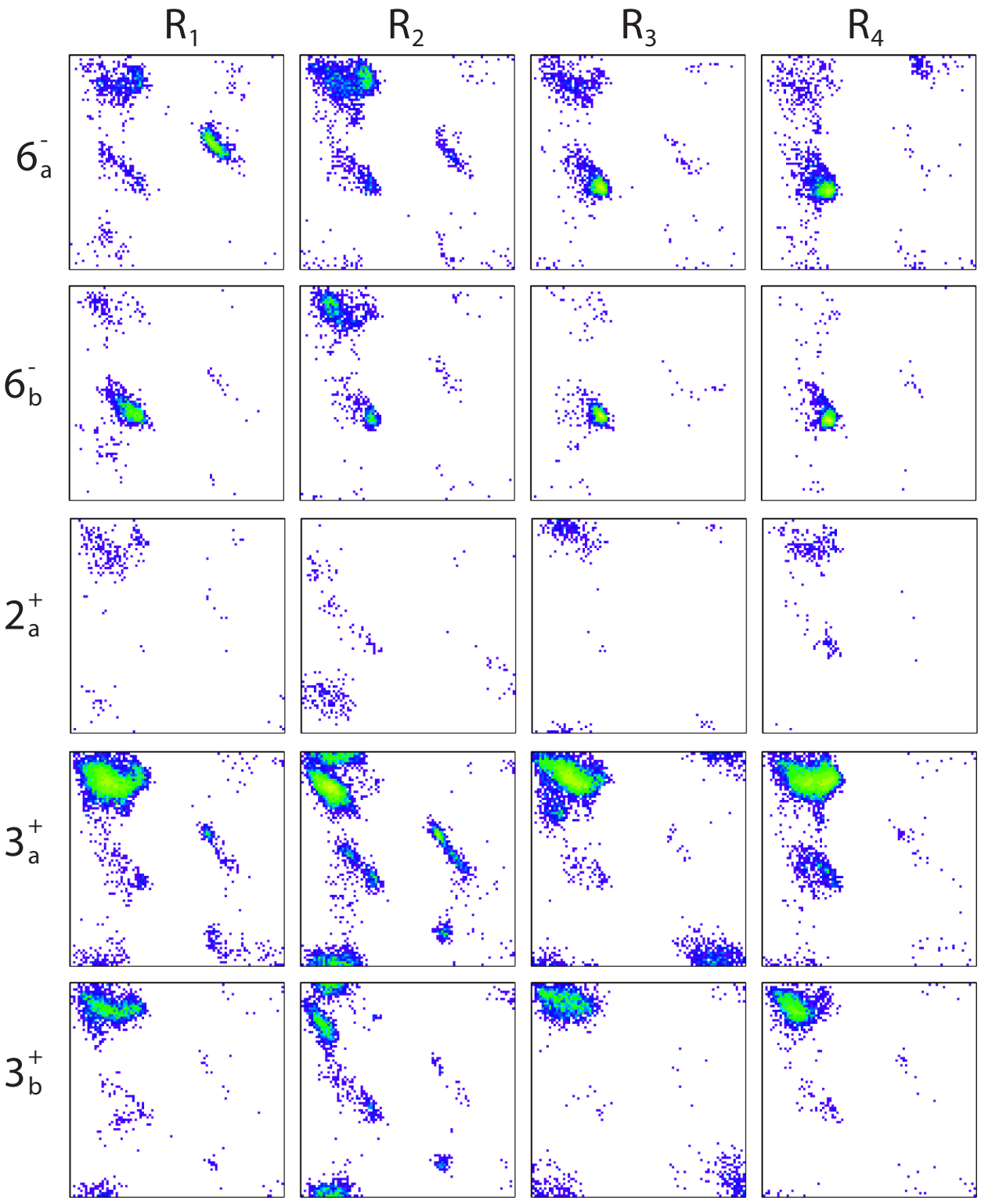
Supplementary Figure 4. Ramachandran plots for the 30 clusters. For each of the 30 clusters in Table 1 Ramachandran plots (Phi,Psi) for position R_1, R_2, R_3 and R_4 of the H-bonds are recorded.



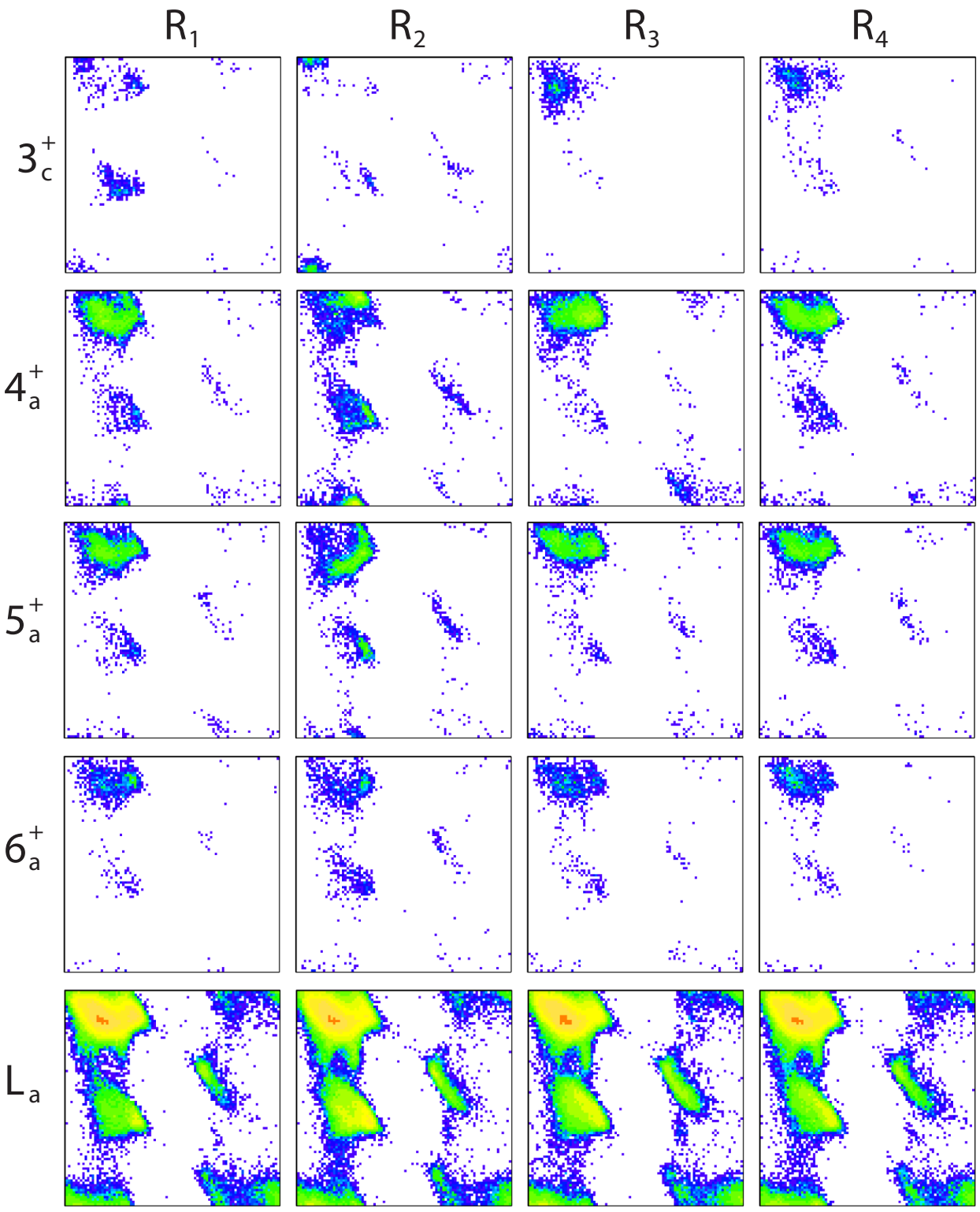
Supplementary Figure 4. (continued).



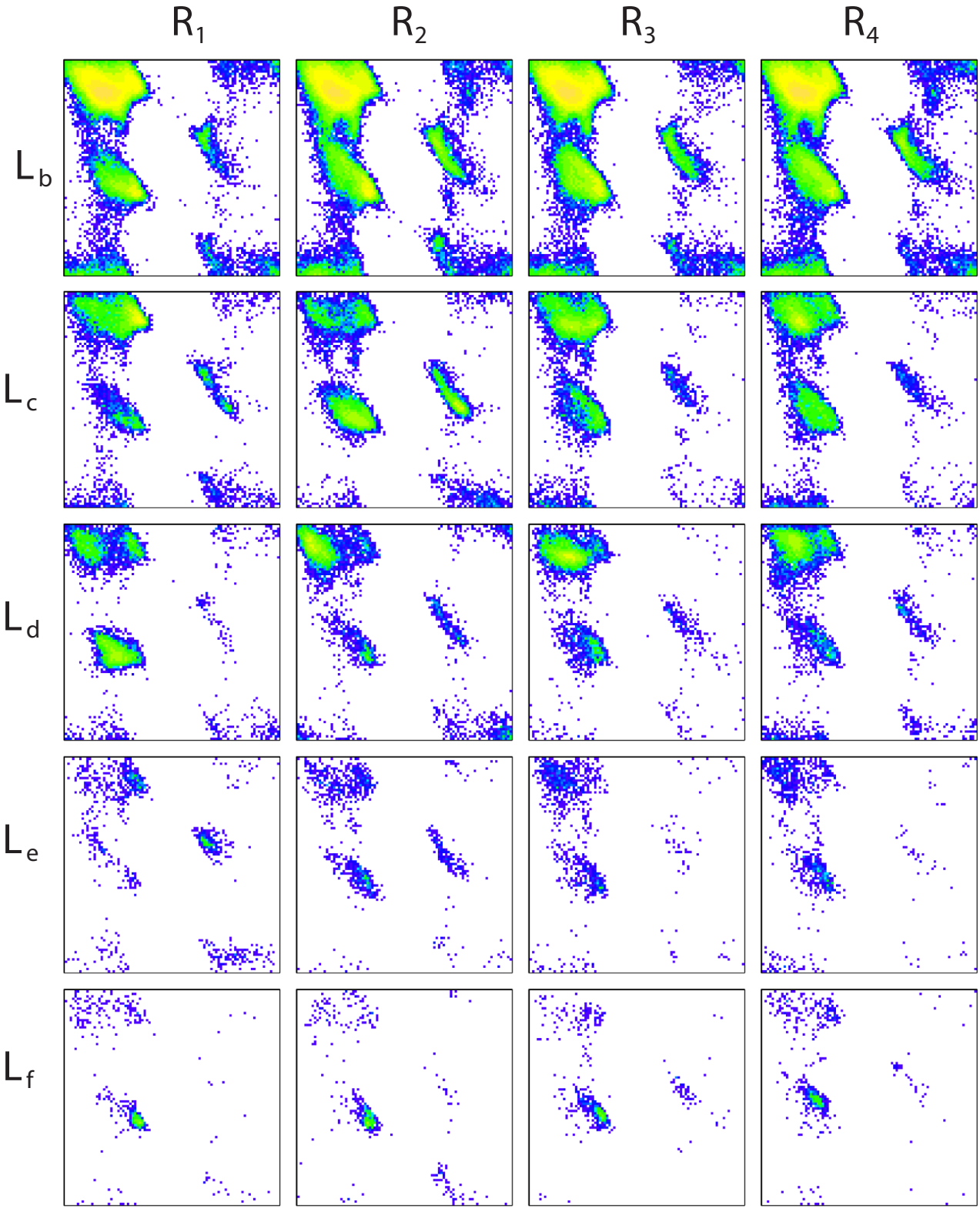
Supplementary Figure 4. (continued).



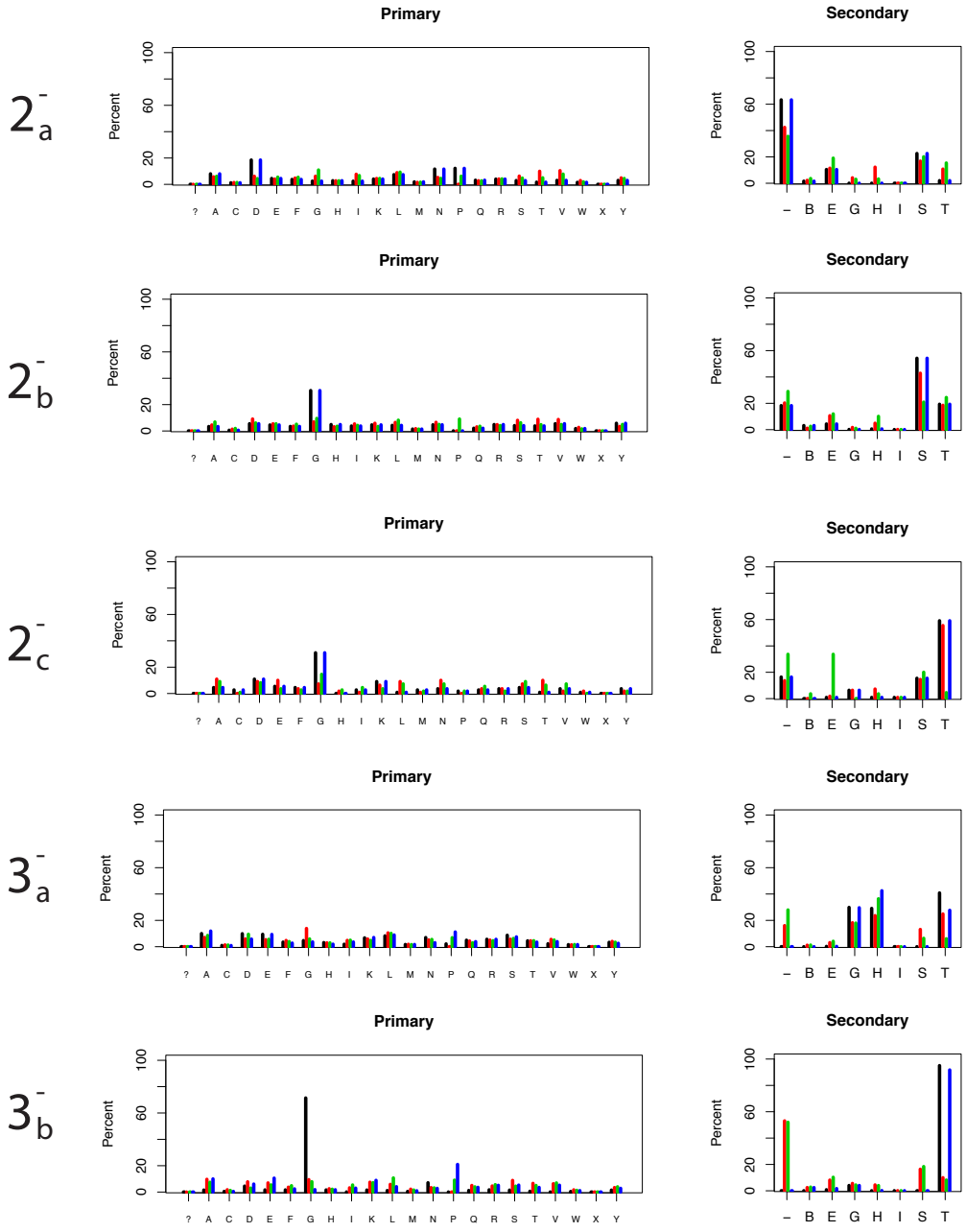
Supplementary Figure 4. (continued).



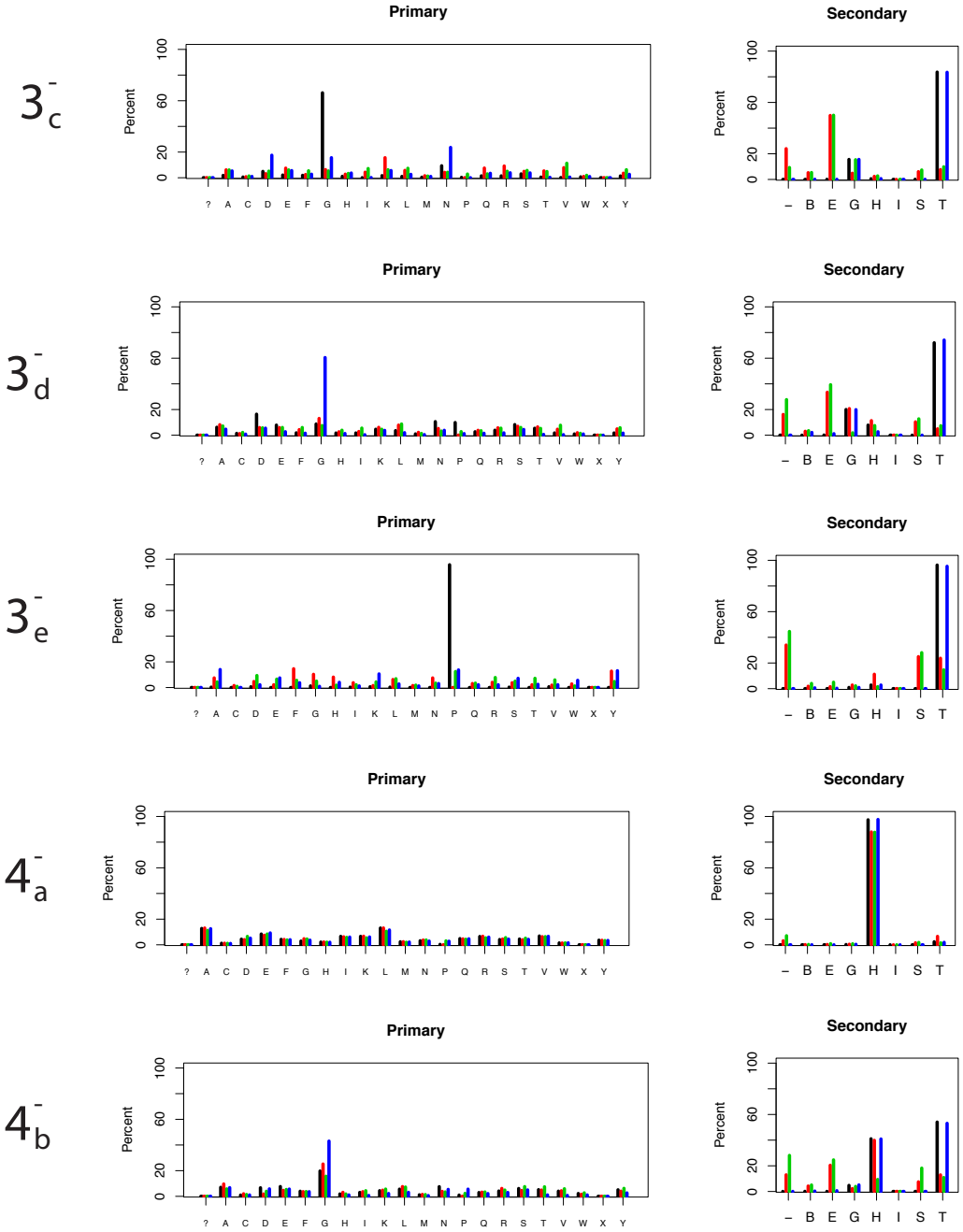
Supplementary Figure 4. (continued).



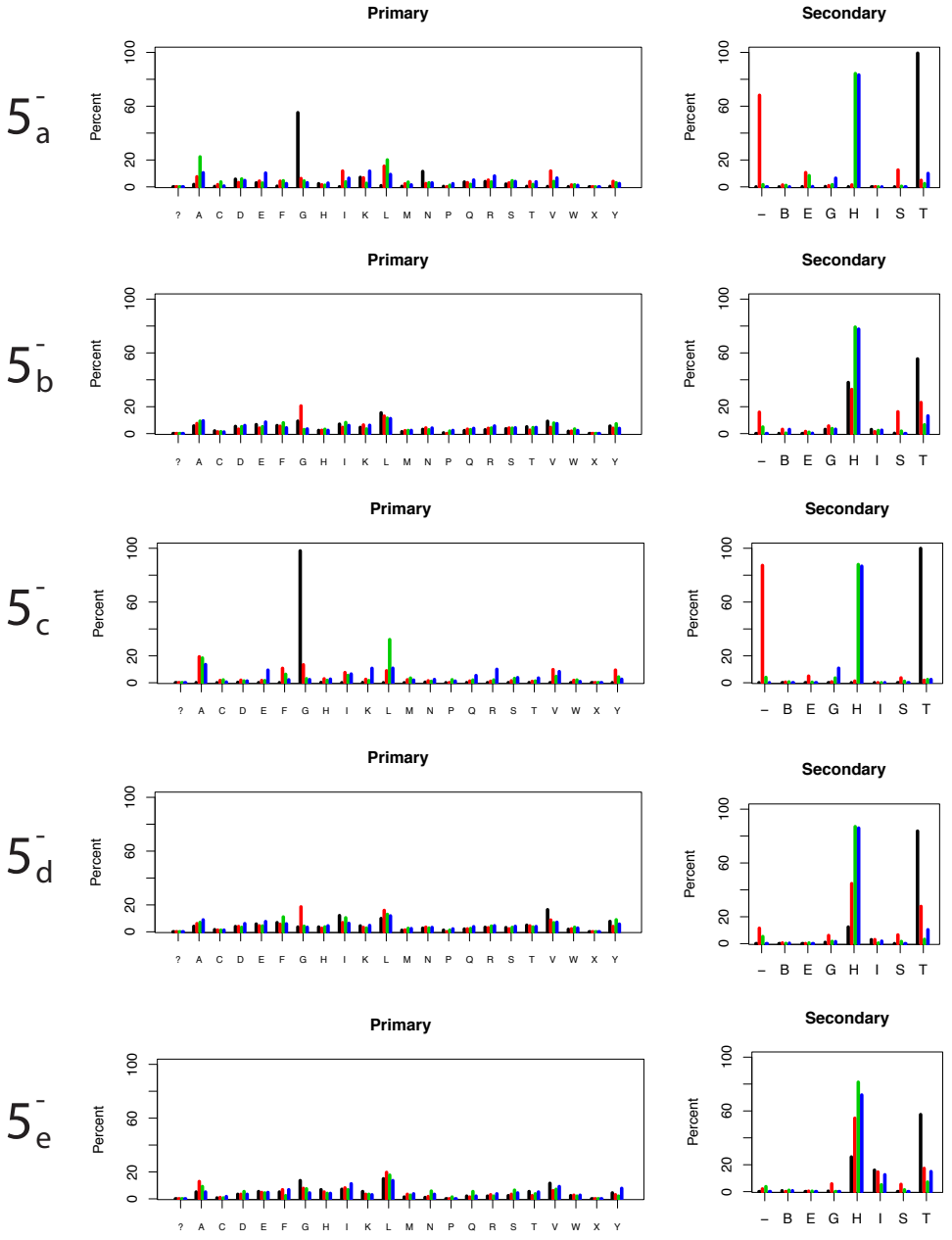
Supplementary Figure 4. (continued).



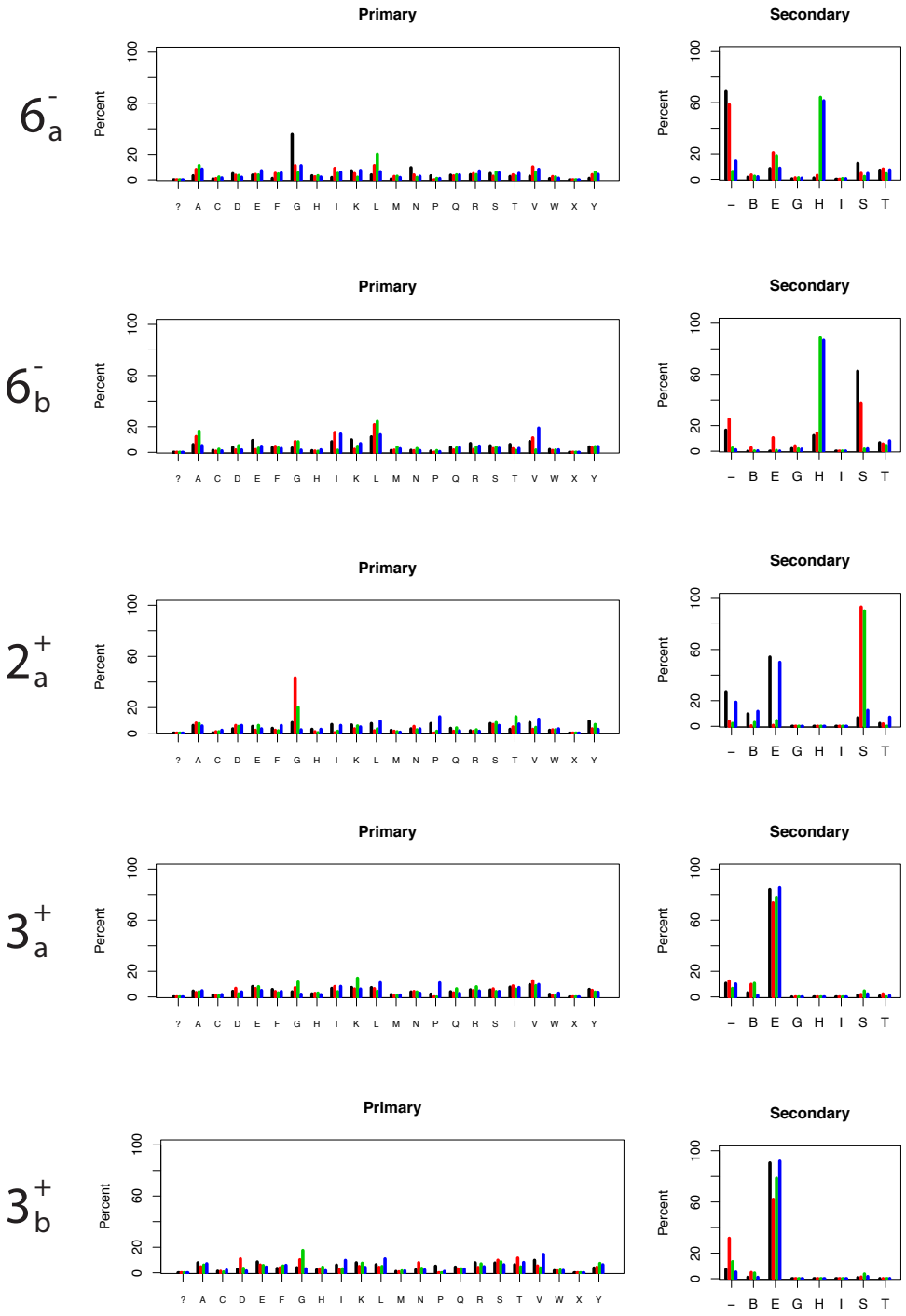
Supplementary Figure 5. Primary and secondary structure plots for the 30 clusters. We use the following special notation. "?" missing data, "X" unspecified or unknown amino acid, "–" Coil



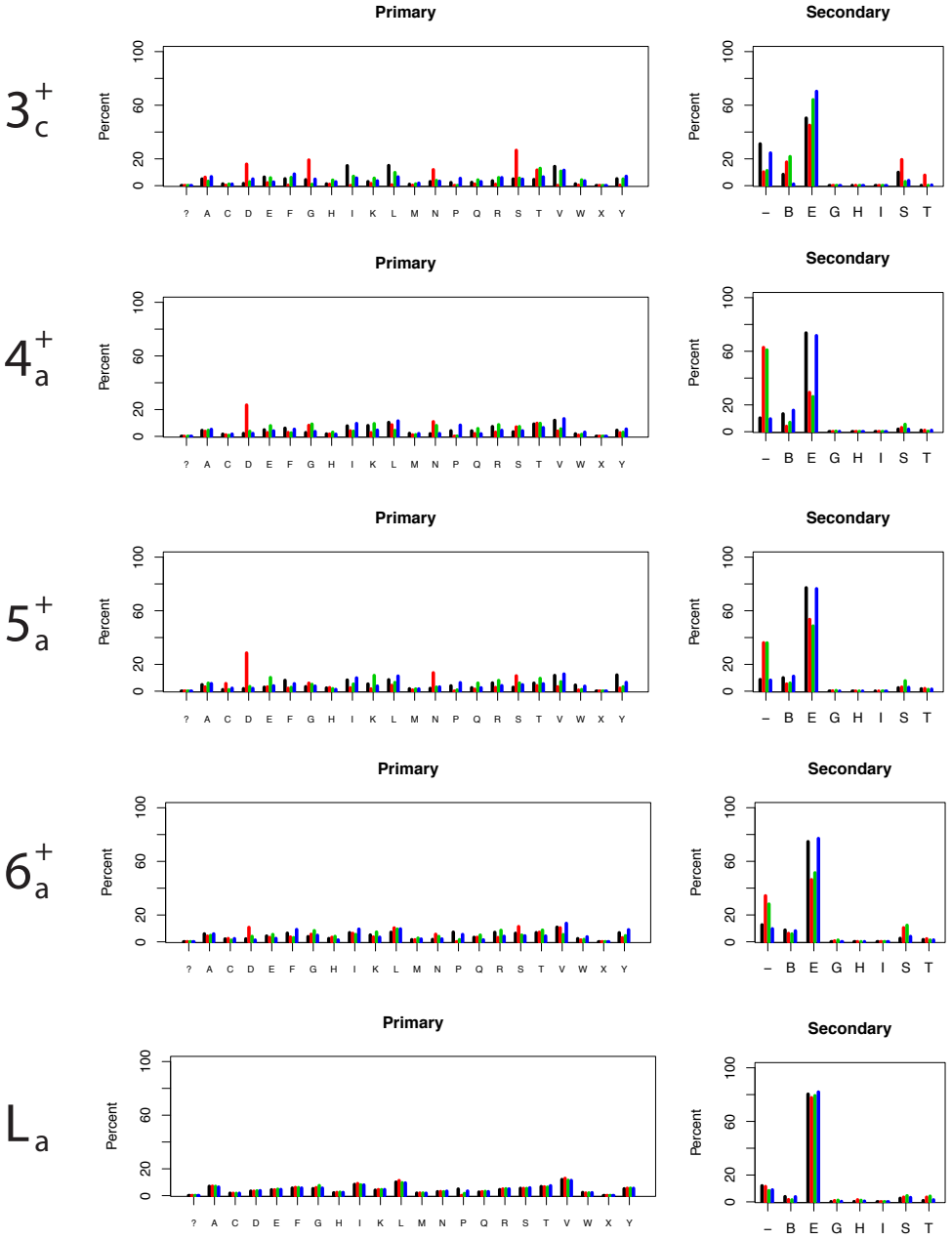
Supplementary Figure 5. (continued).



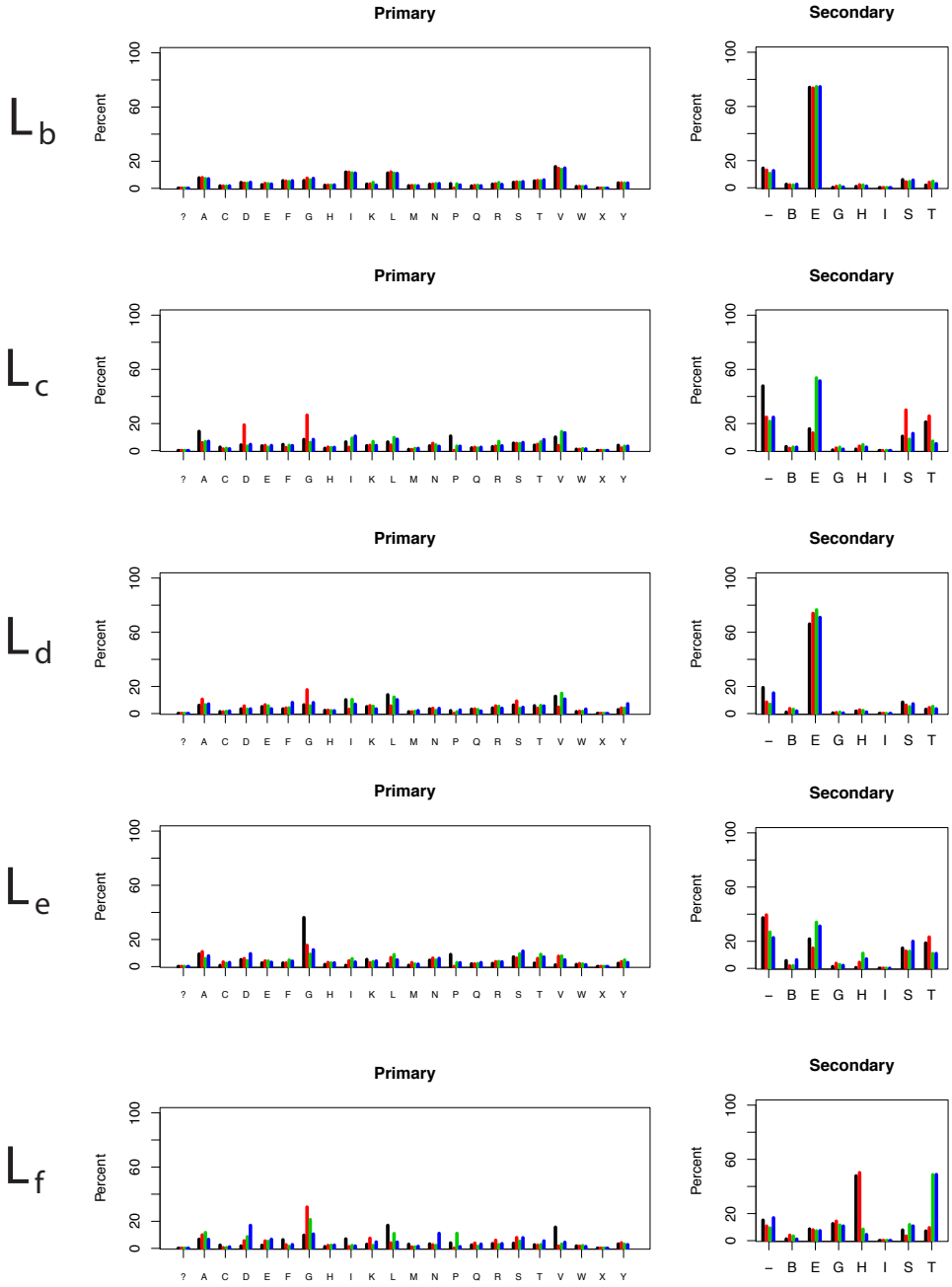
Supplementary Figure 5. (continued).



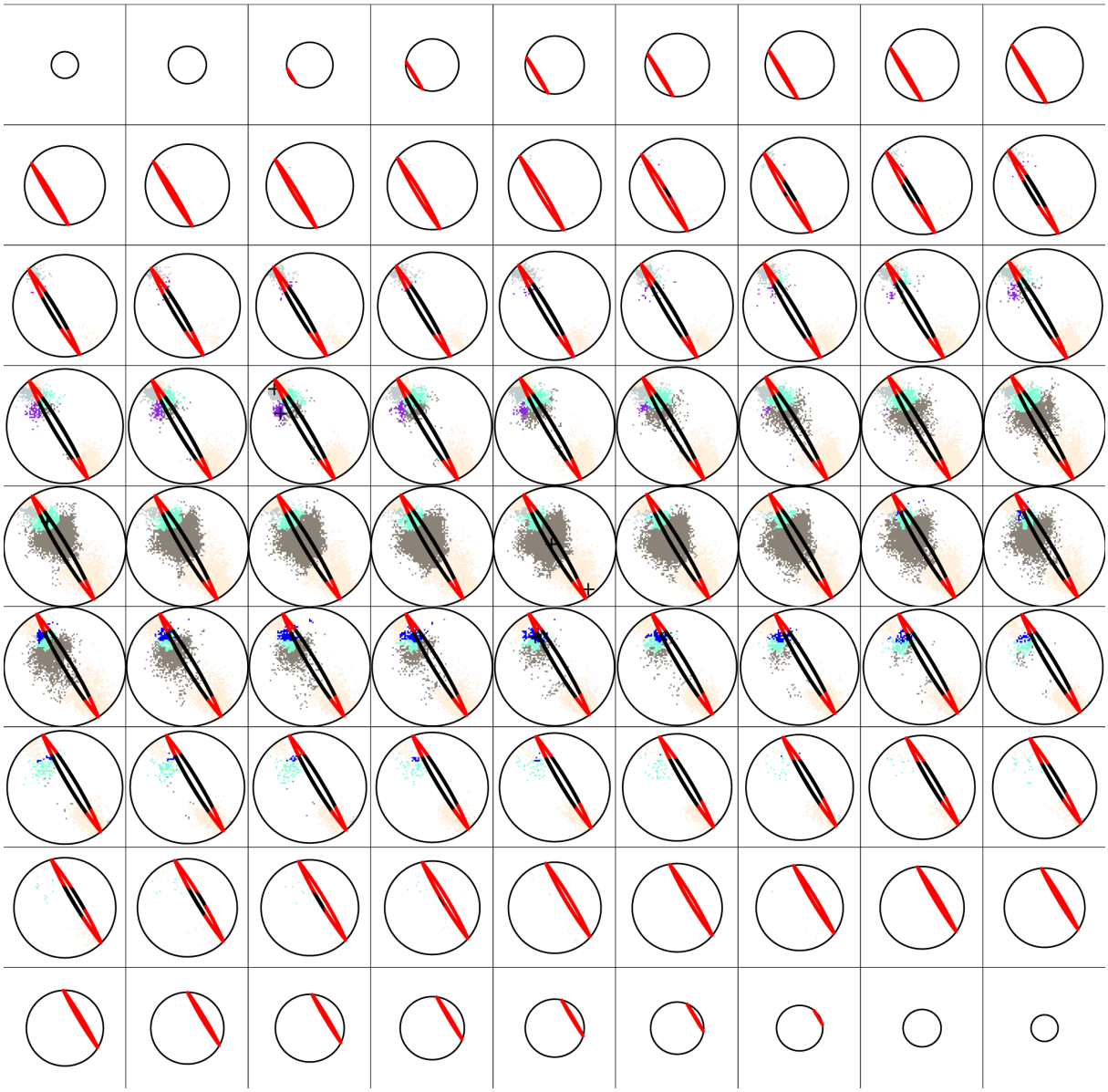
Supplementary Figure 5. (continued).



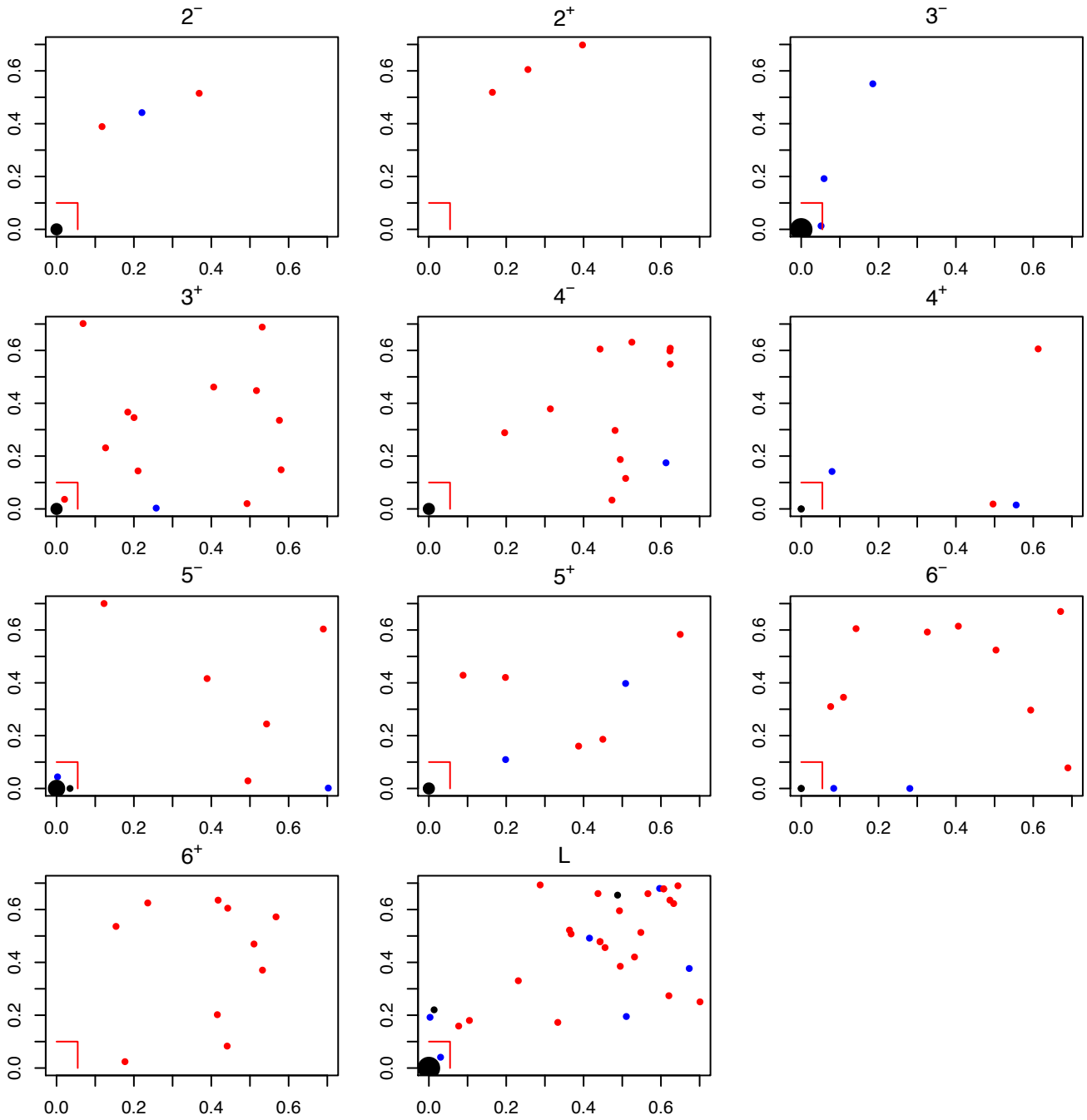
Supplementary Figure 5. (continued).



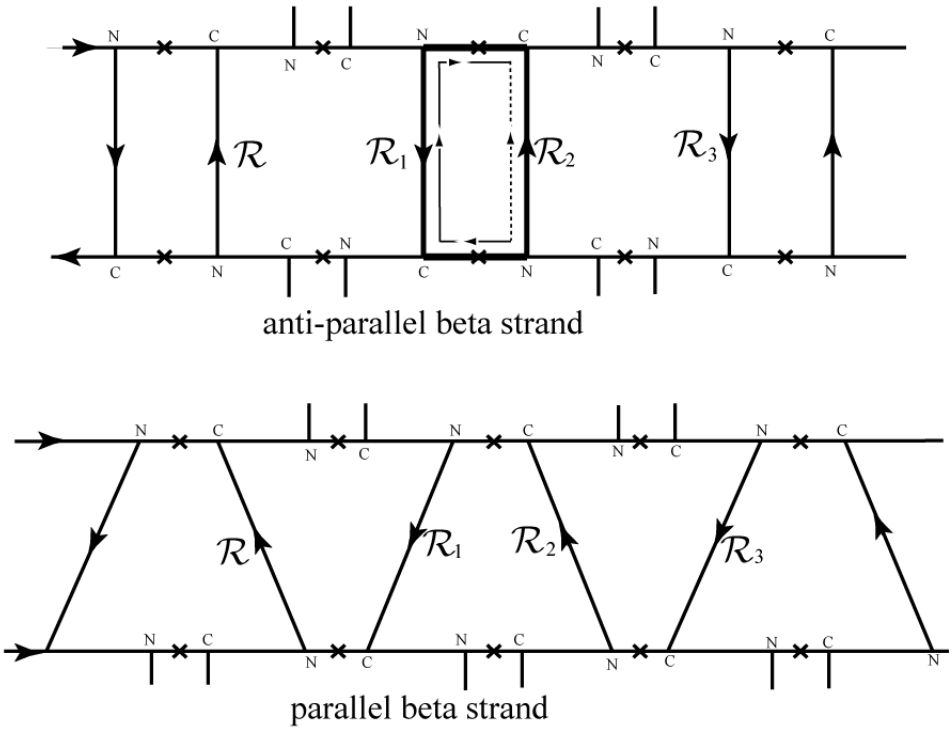
Supplementary Figure 5. (continued).



Supplementary Figure 6. Ideal beta sheet graph in rotational space . Given in Red (part close to the π-sphere) and Black showing the two dimensional surface which forms the set of solutions to the equation $\mathcal{R} = \mathcal{R}_a^B \mathcal{R}^{-1} \mathcal{R}_a^B$. The long range clusters $L_a - L_f$ are indicated in the colours light yellow, dark grey, light green, light grey, dark blue, blue violet respectively.



Supplementary Figure 7. Scatterplot of the squareroot of the two p-values entering the definition of a mode box for the run with a $81 \times 81 \times 81$ grid. The red lines show the limits used in the paper for selecting mode boxes. The point at (0,0) is scaled in size with the number of mode boxes where both p-values are less than 0.0001. The colour of the remaining points reflects the number of Hbonds in the mode box (red: 1-5, blue: 6-10, black: > 10). .



Supplementary Figure 8. Analysis of beta sheets using compositions of rotations along the backbone and across involved H-bonds. Anti-parallel and parallel adjacent beta strands are shown, from which one deduces the equations $\mathcal{R}_1 = \mathcal{R}_a^B \mathcal{R}^{-1} \mathcal{R}_a^B$ for the anti-parallel case and $\mathcal{R}_1 = \mathcal{R}_p^B \mathcal{R}^{-1} (\mathcal{R}_p^B)^{-1}$ for the parallel case.

HQ15

2^-			a	b
	♯		8795	781
		%	100	95
a	16327	98	100/100	0/0
b	1249	85	0/0	100/93
c	110	50	0/0	100/7

HQ30

2^-			a	b
	♯		9821	812
		%	100	100
a	16327	99	100/100	0/0
b	1249	89	0/0	100/91
c	110	67	0/0	100/9

HQ95

2^-			a	b	c
	♯		18971	1460	164
		%	100	97	85
a	16327	100	100/100	0/0	0/0
b	1249	100	0/0	98/100	2/19
c	110	100	0/0	4/0	96/81

HQ60all

2^-			a	b	c	d
	♯		166811	15006	1459	852
		%	5	91	83	24
a	16327	100	23/100	77/100	0/0	0/0
b	1249	100	0/0	0/0	93/100	7/42
c	110	100	0/0	0/0	0/0	100/58

LQ60

2^-			a	b	c
	♯		36935	3077	543
		%	96	76	37
a	16327	100	100/100	0/0	0/2
b	1249	99	0/0	98/99	2/20
c	110	95	0/0	22/1	78/77

CATH

2^-			a	b	c
	♯		22949	2378	200
		%	95	68	22
a	16327	99	100/100	0/0	0/0
b	1249	92	0/0	100/97	0/0
c	110	67	0/0	57/3	43/100

CATHS

2^-			a	b	c	d
	♯		25682	1385	1090	516
		%	94	77	55	21
a	16327	99	100/100	0/0	0/0	0/0
b	1249	91	0/0	70/100	29/98	1/10
c	110	64	0/0	0/0	17/2	83/90

CATHSO

2^-			a	b	c
	♯		34223	3498	324
		%	95	67	20
a	16327	100	100/100	0/0	0/0
b	1249	95	0/0	100/96	0/0
c	110	76	0/0	57/4	43/100

CATHSOL

2^-			a	b	c
	♯		46055	4637	972
		%	94	70	24
a	16327	100	100/100	0/0	0/0
b	1249	96	0/0	99/99	1/11
c	110	82	0/0	12/1	88/89

HQ15

2^+			a
	♯		2
		%	0
a	266	0	NaN/NaN

HQ30

2^+			a
	♯		143
		%	99
a	266	56	100/100

HQ95

2^+			a
	♯		296
		%	89
a	266	97	100/100

HQ60all

2^+			a
	♯		979
		%	27
a	266	100	100/100

LQ60

2^+			a
	♯		560
		%	49
a	266	92	100/100

CATH

2^+			a
	♯		300
		%	41
a	266	45	100/100

CATHS

2^+			a
	♯		322
		%	30
a	266	37	100/100

CATHSO

2^+			a
	♯		439
		%	38
a	266	60	100/100

CATHSOL

2^+			a
	♯		605
		%	30
a	266	63	100/100

Supplementary Table 1. Comparison between the HQ60 clustering and other clusterings Comparison between HQ60 clustering and clustering for other datasets in length category 2^- and 2^+.

HQ15

3^-			a	b	c
	♯		100523	4567	3208
		%	100	99	97
a+b	181593	100	100/100	0/0	0/1
c	7706	95	0/0	100/98	0/1
d	5490	91	0/0	2/1	98/94
e	321	73	0/0	12/0	88/5

HQ30

3^-			a	b	c	d
	♯		111197	4550	2505	923
		%	100	100	100	100
a+b	181593	100	100/100	0/0	0/0	0/0
c	7706	96	0/0	99/98	1/2	0/0
d	5490	94	0/0	1/1	71/92	27/100
e	321	86	4/0	12/1	83/6	1/0

HQ95

3^-			a	b	c
	♯		210275	9228	6813
		%	100	99	98
a+b	181593	100	100/100	0/0	0/0
c	7706	100	0/0	99/99	1/1
d	5490	100	0/0	1/1	99/93
e	321	99	2/0	0/0	98/5

HQ60all

3^-			a	b	c	d	e
	♯		198931	8575	5487	968	272
		%	95	93	99	33	98
a+b	181593	100	100/100	0/0	0/0	0/0	0/0
c	7706	100	0/0	100/100	0/0	0/0	0/0
d	5490	100	0/0	0/0	96/99	4/100	0/0
e	321	100	0/0	0/0	17/1	0/0	83/100

LQ60

3^-			a	b	c
	♯		376953	14216	10936
		%	99	95	92
a+b	181593	100	100/100	0/0	0/0
c	7706	100	0/0	99/98	1/1
d	5490	100	0/0	3/2	97/95
e	321	99	15/0	0/0	85/4

CATH

3^-			a	b	c
	♯		214074	8615	6259
		%	99	94	88
a+b	181593	100	100/100	0/0	0/0
c	7706	98	0/0	100/95	0/1
d	5490	96	0/0	7/5	93/94
e	321	89	5/0	0/0	95/5

CATHS

3^-			a	b	c	d	e
	♯		221982	8261	4556	1568	562
		%	98	91	85	95	80
a+b	181593	100	100/100	0/0	0/0	0/0	0/0
c	7706	98	0/0	99/99	0/1	0/3	0/0
d	5490	95	0/0	2/1	70/99	25/97	3/38
e	321	86	8/0	0/0	2/0	0/0	89/62

CATHSO

3^-			a	b	c
	♯		313765	12127	8805
		%	99	93	90
a+b	181593	100	100/100	0/0	0/0
c	7706	99	0/0	100/98	0/1
d	5490	98	2/0	2/2	96/98
e	321	95	87/0	0/0	13/1

CATHSOL

3^-			a	b	c	d	e
	♯		410917	16235	10662	900	547
		%	99	92	91	81	53
a+b	181593	100	100/100	0/0	0/0	0/0	0/0
c	7706	99	0/0	99/98	1/1	0/0	0/0
d	5490	99	0/0	2/2	89/99	5/57	3/100
e	321	97	42/0	0/0	1/0	57/43	0/0

HQ15

3^+			a	b	c	d
	♯		2318	2081	876	457
		%	97	95	90	77
a	6965	84	46/72	49/100	0/2	5/77
b	2088	80	59/28	0/0	41/67	0/0
c	707	53	0/0	0/0	77/31	22/22

HQ30

3^+			a	b
	♯		3319	2599
		%	100	100
a	6965	89	35/48	65/100
b	2088	88	100/40	0/0
c	707	70	100/12	0/0

HQ95

3^+			a	b	c	d
	♯		8132	1730	937	618
		%	97	98	87	93
a	6965	100	90/92	0/2	9/81	0/1
b	2088	100	27/8	72/97	0/0	1/4
c	707	97	0/0	2/1	23/19	75/95

HQ60all

3^+			a	b	c
	♯		6801	2334	2114
		%	93	95	64
a	6965	100	90/100	0/1	9/49
b	2088	100	0/0	99/95	0/1
c	707	100	0/0	11/4	89/51

LQ60

3^+			a	b	c
	♯		16837	1393	1211
		%	89	59	71
a	6965	99	91/76	9/86	0/1
b	2088	99	98/24	0/0	2/8
c	707	94	3/0	16/13	81/91

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category 3^- and 3^+.

CATH

3^+			a	b	c	d
	♯		5053	3887	2179	797
		%	88	83	77	63
a	6965	90	63/100	32/73	5/20	0/0
b	2088	90	0/0	38/25	60/80	2/6
c	707	67	0/0	10/1	2/1	88/94

CATHS

3^+			a	b
	♯		7753	4304
		%	80	77
a	6965	88	88/98	12/26
b	2088	88	0/0	100/63
c	707	62	25/2	75/11

CATHSO

3^+			a	b	c	d
	♯		11740	2796	1582	915
		%	83	85	73	62
a	6965	93	92/96	3/14	5/41	0/0
b	2088	93	9/2	66/85	25/59	1/5
c	707	79	24/2	2/1	0/0	74/95

CATHSOL

3^+			a	b	c	d	e	f
	♯		14126	4089	1639	1345	1238	134
		%	85	79	38	80	59	33
a	6965	94	86/97	3/11	6/82	4/51	0/1	0/100
b	2088	95	9/3	77/88	0/0	13/49	1/4	0/0
c	707	82	0/0	3/1	17/18	0/0	79/95	0/0

HQ15

4^-			a	b	c
	♯		273923	4613	1088
		%	100	99	83
a	504642	100	98/100	2/100	0/7
b	1996	52	0/0	1/0	99/93

HQ30

4^-			a	b
	♯		313807	1134
		%	100	99
a	504642	100	100/100	0/4
b	1996	66	2/0	98/96

HQ95

4^-			a	b	c
	♯		578308	1226	1176
		%	100	91	83
a	504642	100	100/100	0/1	0/0
b	1996	99	1/0	51/99	47/100

HQ60all

4^-			a	b	c
	♯		541236	1707	1224
		%	99	62	87
a	504642	100	100/100	0/0	0/1
b	1996	100	0/0	50/100	49/99

LQ60

4^-			a	b	c	d
	♯		1115932	2121	1192	610
		%	100	70	51	63
a	504642	100	100/100	0/8	0/0	0/0
b	1996	95	1/0	52/92	30/100	16/100

CATH

4^-			a	b	c
	♯		640102	1219	1094
		%	99	44	65
a	504642	100	100/100	0/0	0/7
b	1996	62	2/0	43/100	55/93

CATHS

4^-			a	b
	♯		655686	2405
		%	99	45
a	504642	100	100/100	0/4
b	1996	55	3/0	97/96

CATHSO

4^-			a	b	c
	♯		923815	1766	1597
		%	99	42	63
a	504642	100	100/100	0/1	0/3
b	1996	74	3/0	44/99	53/97

CATHSOL

4^-			a	b	c	d
	♯		1186274	2340	2174	356
		%	99	32	58	44
a	504642	100	100/100	0/0	0/3	0/0
b	1996	79	4/0	38/100	52/97	7/100

HQ15

4^+			a
	♯		4042
		%	94
a	6848	80	100/100

HQ30

4^+			a
	♯		4256
		%	100
a	6848	87	100/100

HQ95

4^+			a	b
	♯		7051	1080
		%	97	92
a	6848	100	87/100	13/100

HQ60all

4^+			a
	♯		7381
		%	93
a	6848	100	100/100

LQ60

4^+			a
	♯		13384
		%	86
a	6848	99	100/100

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category 3^+ to 4^+.

CATH

4^+			a	b
	♯		6398	1421
		%	84	75
a	6848	86	83/100	17/100

CATHS

4^+			a
	♯		7829
		%	78
a	6848	84	100/100

CATHSO

4^+			a	b
	♯		9916	1261
		%	81	75
a	6848	91	89/100	11/100

CATHSOL

4^+			a	b	c
	♯		12802	1773	691
		%	81	64	86
a	6848	93	84/100	11/100	5/100

HQ15

5^-			a	b	c	d
	♯		10464	2321	955	131
		%	98	93	95	76
a	16501	93	99/84	0/0	0/0	1/100
b	3661	75	1/0	96/85	3/7	0/0
c	3406	91	99/16	1/1	0/0	0/0
d	1907	81	0/0	17/8	83/91	0/0
e	295	76	0/0	88/6	12/2	0/0

HQ30

5^-			a	b
	♯		12015	3609
		%	100	100
a	16501	96	100/84	0/0
b	3661	85	1/0	99/62
c	3406	94	99/16	1/0
d	1907	90	0/0	100/32
e	295	86	0/0	100/5

HQ95

5^-			a	b	c	d	e	f
	♯		16386	6315	4297	2170	384	363
		%	99	99	96	98	85	96
a	16501	100	86/100	12/38	0/0	0/0	2/100	0/0
b	3661	100	1/0	0/0	50/50	49/97	0/0	0/1
c	3406	100	1/0	99/62	0/0	0/0	0/0	0/0
d	1907	100	0/0	0/0	96/50	3/3	0/0	1/7
e	295	100	1/0	0/0	1/0	2/0	0/0	97/93

HQ60all

5^-			a	b	c	d	e	f	g
	♯		18634	7207	3602	1817	1365	436	307
		%	92	72	96	94	29	74	95
a	16501	100	98/99	0/0	0/0	0/0	0/0	2/100	0/0
b	3661	100	0/0	96/94	0/0	1/3	2/56	0/0	0/0
c	3406	100	3/1	0/0	97/100	0/0	0/0	0/0	0/0
d	1907	100	0/0	13/5	0/0	83/97	4/43	0/0	0/0
e	295	100	1/0	0/0	0/0	1/0	0/0	0/0	98/100

LQ60

5^-			a	b	c	d	e
	♯		29448	10979	7554	4083	581
		%	96	94	83	92	56
a	16501	100	85/100	13/40	0/0	0/0	2/100
b	3661	98	0/0	0/0	50/45	50/88	0/0
c	3406	100	1/0	98/60	0/0	0/0	0/0
d	1907	99	0/0	0/0	91/50	9/8	0/0
e	295	98	1/0	0/0	73/5	26/4	0/0

CATH

5^-			a	b	c	d	e
	♯		17102	5928	4061	2516	336
		%	94	93	79	85	46
a	16501	97	88/100	11/34	0/0	0/0	1/100
b	3661	85	0/0	0/0	46/46	54/84	0/0
c	3406	95	1/0	99/66	0/0	0/0	0/0
d	1907	90	0/0	0/0	92/53	8/6	0/0
e	295	82	0/0	0/0	11/1	89/10	0/0

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category 4^+ to 5^-.

CATHS

5^-			a	b	c
	♯		18947	6835	5058
		%	92	78	89
a	16501	96	95/100	0/0	5/20
b	3661	82	0/0	100/60	0/0
c	3406	95	2/0	0/0	98/80
d	1907	88	0/0	100/35	0/0
e	295	76	0/0	100/4	0/0

CATHSO

5^-			a	b	c	d	e	f
	♯		28864	5212	4105	3531	2090	460
		%	94	91	87	83	62	42
a	16501	98	99/97	0/3	0/0	0/0	0/0	1/100
b	3661	90	0/0	0/0	63/88	8/14	29/94	0/0
c	3406	97	16/3	84/97	0/0	0/0	0/1	0/0
d	1907	94	0/0	0/0	6/4	91/85	3/4	0/0
e	295	86	0/0	0/0	87/8	13/1	0/0	0/0

CATHSOL

5^-			a	b	c	d	e	f
	♯		31725	12212	6087	4598	2142	636
		%	93	93	80	83	61	37
a	16501	98	85/100	13/40	0/0	0/0	0/0	1/100
b	3661	92	0/0	0/0	25/32	50/84	24/99	0/0
c	3406	98	1/0	99/60	0/0	0/0	0/1	0/0
d	1907	96	0/0	0/0	92/67	7/6	0/1	0/0
e	295	89	0/0	0/0	17/1	83/10	0/0	0/0

HQ15

5^+			a
	♯		2679
		%	92
a	4532	74	100/100

HQ30

5^+			a
	♯		2796
		%	100
a	4532	83	100/100

HQ95

5^+			a	b	c	d
	♯		3300	1160	731	218
		%	95	93	94	80
a	4532	99	63/100	20/100	14/100	4/100

HQ60all

5^+			a
	♯		4997
		%	91
a	4532	100	100/100

LQ60

5^+			a	b	c
	♯		7169	1001	563
		%	85	68	63
a	4532	98	82/100	12/100	6/100

CATH

5^+			a	b
	♯		4623	435
		%	78	73
a	4532	82	91/100	9/100

CATHS

5^+			a
	♯		5073
		%	72
a	4532	78	100/100

CATHSO

5^+			a	b	c	d	e
	♯		4079	1661	617	468	397
		%	78	79	55	78	47
a	4532	88	58/100	24/100	7/100	7/100	4/100

CATHSOL

5^+			a	b	c
	♯		7552	2405	191
		%	73	71	51
a	4532	90	78/100	21/100	2/100

HQ15

6^-			a	b
	♯		1006	699
		%	84	90
a	1964	55	97/100	3/5
b	1312	64	0/0	100/95

HQ30

6^-			a	b
	♯		1177	808
		%	99	99
a	1964	73	99/99	1/1
b	1312	76	1/1	99/99

HQ95

6^-			a	b
	♯		2312	1458
		%	91	95
a	1964	99	100/100	0/0
b	1312	99	1/0	99/100

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category 5^- to 6^-.

HQ60all

6^-			a	b	c
	♯		3888	2195	836
		%	49	89	10
a	1964	100	6/6	90/100	4/100
b	1312	100	100/94	0/0	0/0

LQ60

6^-			a	b	c
	♯		2944	2188	1463
		%	70	66	75
a	1964	96	4/4	56/100	39/100
b	1312	97	100/96	0/0	0/0

CATH

6^-			a	b	c
	♯		1632	1567	487
		%	59	66	71
a	1964	66	73/100	0/0	27/100
b	1312	75	0/0	100/100	0/0

CATHS

6^-			a	b
	♯		2118	1653
		%	54	62
a	1964	60	98/100	2/2
b	1312	71	0/0	100/98

CATHSO

6^-			a	b
	♯		3105	2335
		%	60	63
a	1964	77	98/100	2/2
b	1312	82	0/0	100/98

CATHSOL

6^-			a	b	c
	♯		3930	3041	54
		%	57	59	22
a	1964	80	97/100	3/3	0/100
b	1312	84	0/0	100/97	0/0

HQ15

6^+			a
	♯		8
		%	0
a	1325	0	NaN/NaN

HQ30

6^+			a
	♯		815
		%	100
a	1325	67	100/100

HQ95

6^+			a
	♯		1498
		%	91
a	1325	98	100/100

HQ60all

6^+			a
	♯		1535
		%	87
a	1325	100	100/100

LQ60

6^+			a
	♯		2967
		%	55
a	1325	94	100/100

CATH

6^+			a
	♯		1655
		%	50
a	1325	58	100/100

CATHS

6^+			a
	♯		1742
		%	40
a	1325	52	100/100

CATHSO

6^+			a
	♯		2484
		%	45
a	1325	70	100/100

CATHSOL

6^+			a
	♯		3214
		%	41
a	1325	75	100/100

HQ15

L			a	b	c	d	e	f
	♯		123961	67881	18755	5768	4320	2087
		%	100	98	99	98	98	98
a	242698	97	85/100	1/3	13/100	1/13	0/0	0/0
b	128347	95	0/0	99/94	0/0	0/0	1/12	0/5
c	13727	89	0/0	16/2	0/0	3/5	54/83	26/89
d	8747	92	4/0	1/0	0/0	93/82	0/1	1/3
e	1221	75	44/0	25/0	0/0	2/0	29/4	0/0
f	808	67	0/0	86/0	0/0	2/0	0/0	12/2

HQ30

L			a	b	c	d	e
	♯		149530	78627	5155	4972	2690
		%	100	100	100	100	100
a	242698	98	99/100	1/3	0/1	0/4	0/2
b	128347	97	0/0	99/95	1/9	0/0	0/1
c	13727	94	0/0	11/1	57/84	1/2	31/90
d	8747	95	9/0	1/0	0/1	86/93	4/7
e	1221	84	27/0	30/0	43/6	0/0	0/0
f	808	80	0/0	92/1	0/0	6/1	2/0

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category 6^+ to L.

HQ95

L			a	b	c	d	e	f	g	h	i
	♯		262711	147587	24626	15299	10728	1986	1984	708	151
		%	100	99	99	99	99	95	98	97	95
a	242698	100	90/100	1/1	9/100	0/0	0/6	0/11	0/1	0/0	0/84
b	128347	100	0/0	98/98	0/0	2/16	0/0	0/3	0/1	0/3	0/16
c	13727	100	0/0	4/0	0/0	80/82	1/2	3/22	12/94	0/0	0/0
d	8747	100	5/0	0/0	0/0	1/1	93/91	0/0	0/2	0/0	0/0
e	1221	100	3/0	1/0	0/0	8/1	0/0	88/64	0/0	0/0	0/0
f	808	100	0/0	5/0	0/0	1/0	11/1	0/0	6/2	77/97	0/0

HQ60all

L			a	b	c	d	e	f	g	h
	♯		259132	150335	17211	15220	1464	897	881	726
		%	95	91	85	95	93	20	81	79
a	242698	100	99/100	0/1	1/13	0/1	0/3	0/31	0/0	0/0
b	128347	100	0/0	99/98	0/0	0/4	0/1	0/0	0/3	0/95
c	13727	100	1/0	4/1	1/1	92/94	2/20	0/24	0/0	0/5
d	8747	100	1/0	0/0	98/85	1/1	0/0	0/45	0/0	0/0
e	1221	100	16/0	3/0	0/0	7/1	75/76	0/0	0/0	0/0
f	808	100	0/0	6/0	12/1	9/1	0/0	0/0	72/97	0/0

LQ60

L			a	b	c	d	e	f	g
	♯		447808	254518	48778	25501	17486	2943	1235
		%	99	97	96	96	96	81	85
a	242698	100	89/100	0/0	10/99	0/1	0/9	0/6	0/0
b	128347	100	0/0	99/98	0/1	1/7	0/0	0/1	0/82
c	13727	100	0/0	12/1	0/0	83/88	0/0	4/30	1/18
d	8747	100	4/0	0/0	0/0	6/3	90/91	0/0	0/0
e	1221	99	14/0	8/0	0/0	6/1	0/0	72/63	0/0
f	808	97	0/0	81/0	0/0	17/1	3/0	0/0	0/0

CATH

L			a	b	c	d	e
	♯		281002	139803	14793	8649	1427
		%	98	97	95	95	77
a	242698	98	99/99	0/1	0/1	0/7	0/93
b	128347	97	0/0	99/97	1/8	0/0	0/7
c	13727	94	0/0	14/1	86/83	0/1	0/0
d	8747	96	14/0	0/0	5/3	81/92	0/0
e	1221	84	9/0	39/0	52/4	0/0	0/0
f	808	77	0/0	78/0	20/1	2/0	0/0

CATHS

L			a	b	c	d
	♯		286598	149501	14008	9615
		%	98	96	95	95
a	242698	98	99/100	0/0	0/1	0/8
b	128347	97	0/0	99/97	1/6	0/0
c	13727	94	0/0	17/2	82/88	0/1
d	8747	95	7/0	0/0	4/2	89/91
e	1221	83	16/0	62/0	22/2	0/0
f	808	72	0/0	53/0	44/2	3/0

CATHSO

L			a	b	c	d	e
	♯		360425	210764	41510	24371	13509
		%	98	97	95	92	95
a	242698	99	88/100	0/0	11/100	0/1	0/5
b	128347	98	0/0	99/99	0/0	1/5	0/0
c	13727	97	0/0	7/1	0/0	91/84	2/3
d	8747	98	7/0	0/0	0/0	2/1	90/92
e	1221	93	9/0	10/0	0/0	81/6	0/0
f	808	84	0/0	4/0	0/0	91/3	4/0

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering for other dataset in length category L.

CATHSOL

L			a	b	c	d	e	f	g	h
	♯		500050	266512	54612	30637	17195	3257	946	213
		%	98	97	95	94	95	75	60	45
a	242698	99	88/100	0/0	11/100	0/1	0/6	0/1	0/18	0/0
b	128347	99	0/0	98/99	0/0	1/8	0/0	0/2	0/82	0/100
c	13727	98	0/0	6/1	0/0	89/83	0/0	5/40	0/0	0/0
d	8747	98	5/0	0/0	0/0	8/5	87/93	0/0	0/0	0/0
e	1221	95	20/0	6/0	0/0	5/0	0/0	68/56	0/0	0/0
f	808	87	0/0	7/0	0/0	91/3	2/0	0/0	0/0	0/0

Grid 51/64

2^-		a	b	c
	♯	16327	1262	97
a	16327	100/100	0/0	0/0
b	1249	0/0	100/99	0/3
c	110	0/0	15/1	85/97

Grid 57/72

2^-		a	b
	♯	16327	1359
a	16327	100/100	0/0
b	1249	0/0	100/92
c	110	0/0	100/8

Grid 71/90

2^-		a	b
	♯	16327	1359
a	16327	100/100	0/0
b	1249	0/0	100/92
c	110	0/0	100/8

Grid 80/100

2^-		a	b
	♯	16325	1358
a	16325	100/100	0/0
b	1248	0/0	100/92
c	110	0/0	100/8

Grid 51/64

2^+		a
	♯	266
a	266	100/100

Grid 57/72

2^+		a
	♯	266
a	266	100/100

Grid 71/90

2^+		a
	♯	266
a	266	100/100

Grid 80/100

2^+		a
	♯	262
a	262	100/100

Grid 51/64

3^-		a	b	c
	♯	181582	7682	5846
a+b	181593	100/100	0/0	0/1
c	7706	0/0	98/98	2/2
d	5490	0/0	2/2	98/92
e	321	2/0	0/0	98/5

Grid 57/72

3^-		a	b	c
	♯	181598	7709	5803
a+b	181593	100/100	0/0	0/0
c	7706	0/0	98/98	1/2
d	5490	0/0	2/2	98/92
e	321	2/0	0/0	98/5

Grid 71/90

3^-		a	b	c	d	e
	♯	100841	80707	7592	5383	586
a+b	181592	56/100	44/100	0/0	0/1	0/1
c	7706	0/0	0/0	98/99	2/3	0/0
d	5490	0/0	0/0	1/1	94/95	5/48
e	321	2/0	0/0	0/0	5/0	93/51

Grid 80/100

3^-		a	b	c	d	e	f
	♯	95439	74955	11172	7764	4913	865
a+b	181591	53/100	41/100	6/100	0/0	0/1	0/0
c	7706	0/0	0/0	0/0	98/97	2/3	0/0
d	5490	0/0	0/0	0/0	3/2	85/95	12/75
e	321	3/0	0/0	0/0	14/1	17/1	67/25

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with different grid size in length category 2^- to 3^-.

Grid 51/64

3^+		a	b	c	d
	♯	8067	703	590	400
a	6965	91/79	8/83	0/2	0/0
b	2088	81/21	0/0	1/3	18/96
c	707	1/0	17/17	80/96	2/4

Grid 57/72

3^+		a	b	c
	♯	8425	768	566
a	6964	91/75	9/82	0/2
b	2088	99/25	0/0	1/2
c	707	4/0	20/18	77/96

Grid 71/90

3^+		a	b	c
	♯	4624	4379	752
a	6961	32/49	58/92	10/88
b	2088	96/43	0/0	4/11
c	706	53/8	46/7	1/1

Grid 80/100

3^+		a	b	c
	♯	5632	2598	1525
a	6961	73/91	10/26	17/76
b	2087	5/2	77/62	18/24
c	707	58/7	42/11	0/0

Grid 51/64

4^-		a	b	c
	♯	504662	1217	732
a	504642	100/100	0/4	0/0
b	1969	4/0	59/96	37/100

Grid 57/72

4^-		a	b
	♯	504694	1913
a	504638	100/100	0/2
b	1969	5/0	95/98

Grid 71/90

4^-		a	b
	♯	504633	1964
a	504631	100/100	0/3
b	1966	3/0	97/97

Grid 80/100

4^-		a	b
	♯	504523	2064
a	504631	100/100	0/7
b	1956	1/0	99/93

Grid 51/64

4^+		a	b	c
	♯	2850	2571	1427
a	6848	42/100	38/100	21/100

Grid 57/72

4^+		a
	♯	6848
a	6848	100/100

Grid 71/90

4^+		a
	♯	6843
a	6843	100/100

Grid 80/100

4^+		a
	♯	6847
a	6847	100/100

Grid 51/64

5^-		a	b	c	d	e
	♯	14651	5614	5015	246	243
a	16500	89/100	0/0	10/33	1/100	0/0
b	3661	0/0	93/61	0/0	0/0	7/100
c	3406	1/0	1/0	99/67	0/0	0/0
d	1907	0/0	100/34	0/0	0/0	0/0
e	295	0/0	100/5	0/0	0/0	0/0

Grid 57/72

5^-		a	b	c
	♯	19905	3640	2223
a	16499	100/83	0/0	0/0
b	3661	1/0	49/49	50/83
c	3406	99/17	1/1	0/0
d	1907	0/0	94/49	5/5
e	295	0/0	4/0	96/13

Grid 71/90

5^-		a	b	c
	♯	14554	5857	5352
a	16499	88/100	0/0	12/36
b	3656	0/0	99/62	1/0
c	3406	0/0	1/0	99/63
d	1907	0/0	100/32	0/0
e	295	0/0	99/5	1/0

Grid 80/100

5^-		a	b	c	d
	♯	10151	5884	5558	4167
a	16495	61/100	0/0	13/40	25/100
b	3657	0/0	100/62	0/0	0/0
c	3406	1/0	1/1	98/60	0/0
d	1907	0/0	100/32	0/0	0/0
e	295	1/0	99/5	0/0	0/0

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with different grid size in length category 3^+ to 5^-.

Grid 51/64

5^+		a
	♯	4525
a	4525	100/100

Grid 57/72

5^+		a	b
	♯	2938	1586
a	4524	65/100	35/100

Grid 71/90

5^+		a	b
	♯	2831	1680
a	4511	63/100	37/100

Grid 80/100

5^+		a	b
	♯	3032	1455
a	4487	68/100	32/100

Grid 51/64

6^-		a	b
	♯	1912	1360
a	1964	97/99	3/5
b	1308	1/1	99/95

Grid 57/72

6^-		a	b
	♯	1977	1289
a	1959	99/99	1/1
b	1307	2/1	98/99

Grid 71/90

6^-		a	b
	♯	1980	1273
a	1952	100/98	0/0
b	1301	2/2	98/100

Grid 80/100

6^-		a	b
	♯	1920	1329
a	1950	97/99	3/4
b	1299	2/1	98/96

Grid 51/64

6^+		a	b
	♯	872	452
a	1324	66/100	34/100

Grid 57/72

6^+		a
	♯	1324
a	1324	100/100

Grid 71/90

6^+		a
	♯	1313
a	1313	100/100

Grid 80/100

6^+		a
	♯	1313
a	1313	100/100

Grid 51/64

L		a	b	c	d	e
	♯	251373	128338	11466	2187	1372
a	242357	100/96	0/0	0/0	0/0	0/8
b	127876	0/0	98/98	1/14	0/0	0/1
c	13727	2/0	11/1	69/82	13/85	4/43
d	8747	96/3	0/0	2/1	2/8	0/0
e	1221	23/0	1/0	22/2	0/0	55/49
f	808	2/0	80/1	0/0	17/6	0/0

Grid 57/72

L		a	b	c	d	e	f	g
	♯	241248	127382	13860	8690	2557	663	336
a	242357	99/100	0/1	0/1	0/11	0/0	0/0	0/15
b	127876	0/0	98/98	1/13	0/0	0/0	0/10	0/85
c	13727	1/0	6/1	77/76	0/1	15/81	0/3	0/0
d	8747	6/0	0/0	3/2	88/88	4/12	0/1	0/0
e	1221	15/0	2/0	83/7	0/0	0/0	0/0	0/0
f	808	0/0	8/0	0/0	1/0	20/6	71/86	0/0

Grid 71/90

L		a	b	c	d	e	f	g	h
	♯	214665	132867	17284	12425	8758	6622	1730	383
a	242357	88/100	1/3	7/100	0/1	0/11	3/100	0/0	0/11
b	127874	0/0	99/95	0/0	1/12	0/0	0/0	0/0	0/89
c	13727	0/0	17/2	0/0	71/78	1/1	0/0	11/87	0/0
d	8747	7/0	0/0	0/0	2/2	88/87	0/0	3/13	0/0
e	1221	14/0	64/1	0/0	22/2	0/0	0/0	0/0	0/0
f	808	0/0	7/0	0/0	91/6	1/0	0/0	0/0	0/0

Grid 80/100

L		a	b	c	d	e	f	g
	♯	216259	136800	13608	9216	8268	7350	3230
a	242356	89/99	2/4	6/100	0/0	0/12	3/100	0/6
b	127874	0/0	99/93	0/0	1/9	0/0	0/0	0/0
c	13727	0/0	20/2	0/0	60/89	0/1	0/0	19/80
d	8747	11/0	1/0	0/0	1/1	82/87	0/0	4/12
e	1221	6/0	90/1	0/0	4/1	0/0	0/0	0/0
f	806	0/0	90/1	0/0	0/0	3/0	0/0	7/2

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with different grid size in length category 5^+ to L.

MS 0.1

2^-	♯	a	b	c	d	s
	16295	1207	69	27	88	
a	16327	100/100	0/0	0/0	0/0	0/36
b	1249	0/0	95/99	0/0	2/78	3/42
c	110	0/0	15/1	63/100	5/22	17/22

MS 0.2

2^-	♯	a	b	s
	16327	1359	0	
a	16327	100/100	0/0	0/NaN
b	1249	0/0	100/92	0/NaN
c	110	0/0	100/8	0/NaN

MS 0.3

2^-	♯	a	b	s
	16327	1359	0	
a	16327	100/100	0/0	0/NaN
b	1249	0/0	100/92	0/NaN
c	110	0/0	100/8	0/NaN

MS 0.1

2^+	♯	a	b	c	d	e	s
	45	35	22	18	16	130	
a	266	17/100	13/100	8/100	7/100	6/100	49/100

MS 0.2

2^+	♯	a	b	c	d	e	s
	125	53	50	20	14	4	
a	266	47/100	20/100	19/100	8/100	5/100	2/100

MS 0.3

2^+	♯	a	s
	266	0	
a	266	100/100	0/NaN

MS 0.1

3^-	♯	a	b	c	d	s
	181582	7688	5193	450	197	
a+b	181593	100/100	0/0	0/0	0/0	0/10
c	7706	0/0	99/99	1/2	0/0	0/3
d	5490	0/0	1/1	93/98	3/32	3/87
e	321	3/0	0/0	1/0	96/68	0/0

MS 0.2

3^-	♯	a	b	c	s
	181637	9040	4433	0	
a+b	181593	100/100	0/0	0/0	0/NaN
c	7706	0/0	100/85	0/0	0/NaN
d	5490	0/0	24/15	75/93	0/NaN
e	321	10/0	1/0	89/6	0/NaN

MS 0.3

3^-	♯	a	b	s
	181715	13395	0	
a+b	181593	100/100	0/0	0/NaN
c	7706	0/0	100/57	0/NaN
d	5490	1/0	99/41	0/NaN
e	321	21/0	79/2	0/NaN

MS 0.1

3^+	♯	a	b	c	d	e	s
	8016	364	181	106	79	1014	
a	6965	89/77	0/1	0/0	0/0	0/19	11/73
b	2088	87/23	1/4	9/100	0/4	0/0	4/8
c	707	1/0	49/94	0/0	14/96	9/81	27/19

MS 0.2

3^+	♯	a	b	c	d	s
	8438	688	403	222	9	
a	6965	91/75	1/6	6/99	2/72	0/100
b	2088	99/24	1/3	0/0	0/0	0/0
c	707	2/0	89/91	1/1	9/28	0/0

MS 0.3

3^+	♯	a	b	c	s
	9179	358	223	0	
a	6965	93/70	5/99	2/73	0/NaN
b	2088	100/23	0/0	0/0	0/NaN
c	707	91/7	1/1	9/27	0/NaN

MS 0.1

4^-	♯	a	b	c	d	e	s
	504583	692	191	136	106	903	
a	504642	100/100	0/0	0/34	0/0	0/0	0/4
b	1969	2/0	35/100	6/66	7/100	5/100	44/96

MS 0.2

4^-	♯	a	b	c	d	e	s
	504692	980	443	183	114	199	
a	504642	100/100	0/3	0/0	0/0	0/0	0/6
b	1969	4/0	48/97	22/100	9/100	6/100	10/94

MS 0.3

4^-	♯	a	b	c	d	s
	504806	885	736	184	0	
a	504642	100/100	0/1	0/0	0/0	0/NaN
b	1969	9/0	44/99	37/100	9/100	0/NaN

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with the Mean shift algorithm in length category 2^- to 4^-.

MS 0.1

4^+		a	s
	♯	6256	592
a	6848	91/100	9/100

MS 0.2

4^+		a	s
	♯	6628	220
a	6848	97/100	3/100

MS 0.3

4^+		a	s
	♯	6782	66
a	6848	99/100	1/100

MS 0.1

5^-		a	b	c	d	e	s
	♯	14722	4630	2099	1880	1519	919
a	16500	89/100	8/28	0/0	0/0	0/0	3/57
b	3661	0/0	0/0	46/81	47/91	0/1	6/24
c	3406	1/0	98/72	0/0	0/0	0/0	1/5
d	1907	0/0	0/0	7/7	9/9	79/99	5/11
e	295	0/0	0/0	89/13	0/0	3/1	8/3

MS 0.2

5^-		a	b	s
	♯	19574	5767	428
a	16500	97/82	0/0	3/97
b	3661	2/0	98/62	0/2
c	3406	100/17	0/0	0/0
d	1907	1/0	99/33	0/0
e	295	1/0	98/5	1/1

MS 0.3

5^-		a	b	s
	♯	19677	5746	346
a	16500	98/82	0/0	2/100
b	3661	2/0	98/62	0/0
c	3406	100/17	0/0	0/0
d	1907	1/0	99/33	0/0
e	295	1/0	99/5	0/0

MS 0.1

5^+		a	b	s
	♯	2361	1248	916
a	4525	52/100	28/100	20/100

MS 0.2

5^+		a	s
	♯	4314	211
a	4525	95/100	5/100

MS 0.3

5^+		a	s
	♯	4346	179
a	4525	96/100	4/100

MS 0.1

6^-		a	b	c	s
	♯	1103	839	391	939
a	1964	0/0	42/99	20/100	38/79
b	1308	84/100	0/1	0/0	15/21

MS 0.2

6^-		a	b	c	d	s
	♯	1327	1184	335	103	323
a	1964	67/99	0/0	17/100	5/100	11/67
b	1308	1/1	91/100	0/0	0/0	8/33

MS 0.3

6^-		a	b	c	d	s
	♯	1338	1235	475	126	98
a	1964	67/99	0/0	24/100	6/100	2/38
b	1308	1/1	94/100	0/0	0/0	5/62

MS 0.1

6^+		a	b	c	d	e	f	s
	♯	188	165	144	127	95	95	511
a	1325	14/100	12/100	11/100	10/100	7/100	7/100	39/100

MS 0.2

6^+		a	b	c	d	e	s
	♯	376	288	262	261	68	70
a	1325	28/100	22/100	20/100	20/100	5/100	5/100

MS 0.3

6^+		a	s
	♯	1318	7
a	1325	99/100	1/100

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with the Mean shift algorithm in length category 4^+ to 6^+.

MS 0.1

L		a	b	c	d	e	f	g	h	s
	♯	237533	125738	14368	8185	656	532	512	430	6785
a	242357	98/100	0/0	0/0	0/9	0/0	0/6	0/0	0/9	2/71
b	127879	0/0	98/99	1/11	0/0	0/5	0/0	0/17	0/5	1/18
c	13727	0/0	5/0	89/85	0/1	0/1	1/25	1/32	0/4	4/8
d	8747	10/0	0/0	4/3	84/90	0/0	0/0	0/0	0/3	1/1
e	1221	13/0	0/0	3/0	0/0	0/0	30/69	21/51	28/78	5/1
f	808	0/0	1/0	12/1	2/0	76/93	0/0	0/0	0/0	9/1

MS 0.2

L		a	b	c	s
	♯	251566	127652	15506	15
a	242357	100/96	0/0	0/1	0/13
b	127879	1/0	98/99	1/8	0/27
c	13727	2/0	10/1	88/78	0/0
d	8747	97/3	0/0	3/2	0/0
e	1221	14/0	0/0	85/7	1/60
f	808	5/0	4/0	91/5	0/0

MS 0.3

L		a	b	s
	♯	251701	143038	0
a	242357	100/96	0/0	0/NaN
b	127879	0/0	100/89	0/NaN
c	13727	3/0	97/9	0/NaN
d	8747	98/3	2/0	0/NaN
e	1221	16/0	84/1	0/NaN
f	808	10/0	90/1	0/NaN

Supplementary Table 1 (Continued). Comparison between the HQ60 clustering and clustering with the Mean shift algorithm in length category L.

2^-_a	PVLP(0.37)	PILP(0.37)	PLLP(0.28)	DLLD(0.28)	DLGD(0.25)	DVLD(0.24)	DVGD(0.23)	PVVP(0.21)	DFGD(0.20)	NNVN(0.19)
2^-_b	YNYY(0.80)	GDLG(0.64)	GSTG(0.48)	GSLG(0.48)	GEEG(0.48)	GGVG(0.40)	GKIG(0.40)	GLGG(0.40)	GIGG(0.40)	GVFG(0.40)
2^-_c	KLGK(1.82)	KAAK(1.82)	GDVG(1.82)	GTSG(1.82)	GDKG(1.82)	YKRY(0.91)	YAGY(0.91)	YNFY(0.91)	YYDY(0.91)	WLPW(0.91)
2^+_a	IGGV(0.75)	LGDV(0.75)	AGQT(0.75)	MGTP(0.75)	PGSP(0.75)	SMNP(0.75)	HKGP(0.75)	GKGP(0.75)	LGGL(0.75)	ASGH(0.75)
3^-_a	AGAA(0.04)	LGKE(0.03)	AGLA(0.03)	AGKA(0.03)	AGAE(0.03)	AGLE(0.03)	NGDE(0.03)	SLDP(0.03)	LGKA(0.03)	AGEA(0.03)
3^-_b	GDKP(0.26)	GKLK(0.22)	GGDS(0.20)	GSAP(0.20)	GSPA(0.19)	GDRP(0.18)	GAAP(0.18)	GTPA(0.18)	GTAP(0.17)	GDKK(0.17)
3^-_c	GKVN(0.65)	GKKD(0.64)	GKVD(0.48)	GKID(0.40)	GVVN(0.34)	GVVD(0.34)	GKEN(0.32)	GKKN(0.31)	GQVD(0.31)	GEVD(0.27)
3^-_d	LGPG(0.29)	LGHS(0.24)	DLLG(0.22)	DKKG(0.22)	DALG(0.20)	MGHS(0.18)	DKEG(0.18)	SGSG(0.16)	DKLG(0.16)	GGLG(0.16)
3^-_e	PFGY(0.93)	PAPS(0.93)	PARK(0.93)	PGVE(0.93)	PYLE(0.93)	PHLE(0.93)	PGVY(0.62)	PQSY(0.62)	PQRY(0.62)	PYGY(0.62)
3^+_a	YTGN(0.11)	VKKI(0.10)	KTGY(0.09)	ETGV(0.09)	KSNP(0.09)	EVKK(0.09)	EGGK(0.09)	VTGV(0.07)	LGGV(0.07)	GEYI(0.07)
3^+_b	QTGV(0.14)	ITGV(0.14)	ADGV(0.14)	TLLT(0.14)	CGHT(0.14)	AGHT(0.14)	AGHS(0.14)	VTGC(0.14)	LSSY(0.10)	EDSY(0.10)
3^+_c	VDTV(0.57)	VDTL(0.57)	VGLD(0.57)	IDTV(0.42)	FALV(0.42)	GKMP(0.42)	VSEA(0.42)	ISVY(0.28)	INVY(0.28)	ADHY(0.28)
4^-_a	AAAA(0.05)	AAAL(0.04)	LAAA(0.04)	ALLA(0.04)	ALAA(0.03)	EAAL(0.03)	AALA(0.03)	ALAL(0.03)	LAAL(0.03)	ALLL(0.03)
4^-_b	GRRG(0.66)	GLSG(0.51)	YCYG(0.41)	GLEE(0.41)	EGLG(0.36)	GRSG(0.30)	AERG(0.30)	TGGG(0.30)	AGGG(0.30)	GTGF(0.30)
4^+_a	LIWV(0.19)	VDKP(0.19)	FEGY(0.16)	FDNL(0.16)	VDNV(0.15)	LDNV(0.15)	IDNV(0.13)	LDNP(0.13)	LDNL(0.13)	VDTV(0.12)
5^-_a	GILK(0.44)	GLAK(0.33)	GVLE(0.32)	GLLE(0.30)	GLAE(0.30)	GLLK(0.28)	GVLK(0.27)	GVAK(0.27)	GVLR(0.27)	GLAR(0.27)
5^-_b	EAQM(0.16)	LLVR(0.14)	LGVE(0.14)	THPA(0.14)	VLFY(0.11)	VLFV(0.11)	LGEV(0.11)	DGAV(0.11)	KGDN(0.11)	GGPL(0.11)
5^-_c	GAAL(1.17)	GALA(1.00)	GALK(0.88)	GALL(0.82)	GVLA(0.76)	GYLK(0.73)	GALR(0.68)	GYLA(0.68)	GAAV(0.65)	GAAI(0.65)
5^-_d	KHLH(0.31)	SEIQ(0.26)	HGEW(0.21)	FALV(0.21)	KHFH(0.21)	VEYF(0.21)	VLFE(0.21)	HHHY(0.16)	VLFY(0.16)	YLVV(0.16)
5^-_e	HHHH(1.69)	HHHY(1.02)	AALL(1.02)	TLDL(0.68)	GHHE(0.68)	LGYY(0.34)	QNVY(0.34)	WGVY(0.34)	IVTY(0.34)	ALTY(0.34)
5^+_a	IDKV(0.24)	VDKV(0.20)	YDAY(0.13)	VDEV(0.13)	LDKL(0.13)	IDKL(0.13)	YDEY(0.11)	VDTV(0.11)	IDTV(0.11)	VDSV(0.11)
6^-_a	GVLR(0.41)	GLLR(0.36)	GILA(0.36)	SDAA(0.36)	GILE(0.31)	GLLA(0.31)	GILK(0.25)	GVLF(0.25)	GLLV(0.20)	GLLT(0.20)
6^-_b	IILK(0.46)	LALV(0.38)	CVCV(0.38)	RIAV(0.31)	ILLK(0.31)	ELRV(0.23)	RRQV(0.23)	KVLV(0.23)	VLLV(0.23)	LLLV(0.23)
6^+_a	PSRR(0.23)	RVLR(0.23)	RLTL(0.23)	RDRF(0.23)	GFMF(0.23)	TQNY(0.15)	RVLY(0.15)	RLLY(0.15)	FEKY(0.15)	ESFY(0.15)
L_a	VLVV(0.04)	VVVV(0.03)	VKEV(0.03)	VERV(0.03)	LVVL(0.03)	VTTV(0.03)	TLLT(0.03)	VLLV(0.02)	VVLV(0.02)	TLVT(0.02)
L_a	VVVV(0.10)	VIVV(0.09)	VLVV(0.09)	VVLV(0.09)	VIVI(0.08)	VILV(0.08)	VLLV(0.08)	VLVI(0.07)	VVIV(0.07)	VVVL(0.07)
L_a	ADVV(0.26)	ADKV(0.21)	ADKI(0.18)	ADVI(0.17)	SGVS(0.16)	PDKI(0.16)	VDVV(0.15)	YDKV(0.15)	PGVT(0.15)	VDKI(0.15)
L_a	IGVV(0.10)	LGVV(0.10)	LALV(0.10)	VGIV(0.10)	IGVF(0.10)	VGLV(0.09)	LGVY(0.08)	VGLY(0.08)	LLLV(0.08)	LGVR(0.08)
L_a	GAVD(0.57)	KSTS(0.41)	WGKN(0.41)	KVTS(0.33)	GLSS(0.33)	GALE(0.33)	GSGV(0.25)	GCYT(0.25)	DNYS(0.25)	KITS(0.25)
L_a	VRES(0.50)	VGDN(0.50)	FDGG(0.50)	LGPD(0.50)	GGDY(0.37)	LFGG(0.37)	VKLE(0.37)	VRLD(0.37)	LRLD(0.37)	VKDD(0.37)

Supplementary Table 2. Most frequent flanking primary four tubles. For each cluster the 10 most frequent four-tubles of primary structures are listed. The percentage occurence within the cluster of this 4-tuble is given in parenthesis.

Structure	Backbone angles	H-bond rotation	Cluster containing this rotation
Classic gamma	(70,-60)	2.999(0.196,0.829,0.525)	2_b^-
Inverse gamma	(-70,60)	2.999(-0.196,-0.829,0.525)	2_a^-
Beta I	(-60,-30), (-90,0)	2.556(-0.378,0.822,-0.426)	3_a^-
Beta I'	(60,30), (90,0)	2.556(0.378,-0.822,-0.426)	3_c^-
Beta II	(-60,120), (80,0)	3.114(0.511,-0.724,0.464)	3_b^-
Beta II'	(60,-120), (-80,0)	3.114(-0.511,0.724,0.464)	3_d^-
Beta VIa1	(-60,120), (-90,0)	2.993(0.455,-0.711,-0.536)	3_d^-
Beta VIa2	(-120,120), (-60,0)	3.142(0.682,-0.727,-0.079)	(3_d^-)
Beta VIb	(-135,135), (-75,160)	2.488(0.001,0.357,-0.934)	(3_a^-)
Beta VIII	(-60,-30), (-120,120)	2.351(0.081,0.140,-0.987)	(3_a^-)
3_{10} helix	(-70,-20)	2.462(-0.201,0.933,-0.299)	3_a^-
Right alpha helix	(-65,-40)	1.086(-0.315,0.935,-0.164)	4_a^-
Left alpha helix	(60,60)	1.571(0.418,-0.908,0.000)	(4_b^-)
Alpha IRS	(-60,-29), (-72,-29), (-96,-20)	1.111(-0.673,0.732,-0.103)	4_a^-
Alpha ILS	(48,42), (67,33), (70,32)	0.891(0.628,-0.752,-0.200)	4_b^-
Alpha IIRS	(-59,129), (88,-16), (-91,-32)	1.515(-0.550,0.694,-0.465)	4_a^-
Alpha IILS	(53,-137), (-95,81), (57,38)	1.665(0.721,-0.691,-0.059)	4_b^-
Alpha IRU	(59,-157), (-67,-29), (-68,-39)	2.593(-0.847,0.449,0.284)	4_b^-
Alpha ILU	(-61,158), (64,37), (62,39)	2.601(0.850,-0.425,0.311)	4_b^-
Alpha IIRU	(54,39), (67,-5), (-125,-34)	2.809(0.823,-0.522,-0.222)	4_b^-
Alpha IILU	(-65,-20), (-90,16), (86,37)	2.972(0.772,-0.566,0.290)	4_b^-
Pi helix	(-55,-70)	0.196(-0.446,0.889,0.109)	5_b^-
HB-AAAa	(-66,-33), (-71,-31), (-97,-6), (65,38)	2.780(-0.631,0.738,0.241)	5_a^-
HB-PgAA	(-62,130), (85,-8), (-113,-64), (-101,-18)	0.837(-1.000,-0.003,0.029)	5_b^-
HB-AAAA	(-63,-24), (-87,-15), (-115,-71), (-99,-18)	1.082(-0.803,0.003,0.596)	5_b^-
Schellman	(-65,-42), (-63,-30), (-90,4), (76,24)	2.394(-0.655,0.675,0.339)	5_a^-

Supplementary Table 3. Rotations as defined through backbone conformational angles and the clusters they belong to. Clusters in parentices indicate not belonging to any cluster, but closest to the indicated cluster.

Cluster	Total	3_{10}+I	I'	II	II'	VIa1	VIa2	VIb	VIII	Non
3_a^-	158784	144550	20	143	16	306	164	311	1490	11784
3_b^-	16109	251	51	13313	16	38	26	57	17	2340
3_c^-	7447	45	5532	6	181	24	5	80	15	1559
3_d^-	5315	244	132	10	3920	303	3	19	90	594
3_e^-	313	12	0	1	4	254	1	0	1	40

Supplementary Table 4. Number of occurrences of different Beta turns Number of occurrences of different Beta turns within the five clusters $3_a^-, 3_b^-, 3_c^-, 3_d^-$ and 3_e^-. The classification is based on conformational angles when available.

Cluster	Both sides	Only before	Only after
3_a^-	4%	18%	17.5%
4_a^-	65.5%	6.7%	10.8%
5_b^-	2%	25%	4.5%
5_d^-	2%	3%	48%
5_e^-	37%	28%	22%

Supplementary Table 5. Fraction of cluster in helixes. Proportion of cluster which has a parallel H-bond of the same length category on both sides, immediately after only and immediately before only.

Anti-parallel	(-135,150)	$\mathcal{R}_a^B$=2.928(-0.806,-0.475,0.353)
Parallel	(-120,135)	$\mathcal{R}_p^B$=2.928(-0.713,-0.528,0.461)

Supplementary Table 6. Ideal beta backbone conformational angles and rotations.

Supplementary Note 1. Database specifications.

The data sets was derived using two types of services. The PISCES procedure chooses subsets from the full PDB data base and allows for specification of a threshold for a desired quality and a maximum sequence homology. Here we use only X-ray structures. HQ data sets were derived using the PISCES cull server with a resolution threshold of 2.0Å and Rfac threshold of 0.2. This way we obtained the dataset HQ15, HQ30, HQ60 and HQ95 with maximum sequence homology $15\%, 30\%, 60\%$ and 95% respectively. The H-bonds which are compiled in these dataset are all DSSP H-bonds which satisfy the further condition (3). In order to investigate the significance of this extra condition, we also compiled the set HQ60all, which just takes all DSSP bonds from the high quality structures at maximum sequence homology of 60%.

Further the datasets LQ15, LQ30, LQ60, LQ95 were derived by requiring only Res < 3.0Å and Rfac < 0.3, at the maximum sequence homology of $15\%, 30\%, 60\%, 95\%$.

We note that the datasets HQ15, LQ15, HQ30, LQ30, HQ60, LQ60, HQ95, LQ95, HQ60all can possibly contain repeated domains, however this does not pose a real problem due to the rare occurrence of this in our main standard database. For HQ60 there are 5784 unique domains from CATHS and there are 876 pairs, 187 triples, 62 quadruples, 19 quintuples, 11 sixtuples og 3 seventuples. If one considers the domains at the CATHSO level all domains are uniquely represented except for 56 pairs and at the CATHSOL level it is only 7 pairs and no higher tuples in any of these two. These numbers of possible duplications of bonds pose no problem in terms of the clustering, as is also documented by the comparison of all clustering runs below in Supplementary Note 4.

The CATH data set was downloaded from the newest version of the CATH database (version 4.0.0). CATHS, CATHSO, CATHSOL, denote datasets defined by a maximum sequence homology of 35%, 60% and 95%, respectively derived directly from the CATH database. In addition another dataset was used, namely the following one generated by the CATH developers containing non-redundant domains. It is a non-redundant subset of 16,983 of the CATH domains that contains no pair of domains that (according to BLAST) shares $\geq 40\%$ sequence identity over $\geq 60\%$ overlap (over the longer sequence) and is as big as possible a subset, under these conditions, from the S100 chain set of the CATH database. We here simply denote this dataset CATH.

Supplementary Note 2. Positioning of the H-atoms at the amid end of the peptide units.

Backbone amide H atoms were placed based on statistics on known H coordinates in a protein structure, PDB ID 1CEX, solved to atomic resolution ($R = 1.0$Å). For this structure the 3-frames were generated as described in the Method section for each peptide plane. For each 3-frame the local coordinates, within the frame, of the HN atoms were calculated. It was observed that the x, and y-coordinates had very little variation (approximately 0.01Å) whereas the z-coordinate had a larger but still small variation of 0.1Å. Since the H atoms this way are inferred, we also check the clustering up against the clustering for the dataset HQ60all, where condition (3) is not use, and hence the position of the H atoms in this condition does not play any role. The result is discussed in Supplementary Note 4 below.

Supplementary Note 3. Selection and adaption of clustering algorithms.

Our main clustering algorithm is based on defining clusters in terms of seed points being sufficiently strong modes of the density. The number of clusters is therefore defined by the number of significant local maxima of the density function. This approach requires the estimation of the density and a procedure for finding the local maxima. Having established the seed points, data need to be assigned to one of the seeds or become unclassified. Please see the Method section for further details on our main clustering algorithm.

In the literature this approach is known under the heading of 'mode seeking',[14,15] where modes of a mixture density is found by ascending iterations starting from any initial point. These methods are designed for euclidean space and cannot be directly implemented for our data in rotational space. The *Mean shift* algorithm has, however, recently been developed for more general spaces[16]. This has therefore allowed us to implement this algorithm as well. In the following Supplementary Note 4 we provide a comparison between runs form the two algorithms. The Mean shift algorithm has a smoothing parameter and we run the algorithm with three choices of this parameter.

We also remark that we need implementations that can handle large dataset (of the order 300000 points).

Other approaches to clustering, such as for example DBSCAN, are not based on finding modes, but rely on density-connectedness. In such an approach two highly significant modes can be joined into the same cluster due to spacial contact of their corresponding clusters.

Supplementary Note 4. Comparison of clustering runs.

A rather comprehensive study of the stability of the clustering of the data has been undertaken with the view to uncover how stable our clustering is under changes of the input dataset detailed above, adjustment of parameters in the clustering algorithms and further the clustering algorithm itself.

For the comparison of clusters from two dataset, we have (re-)fused cluster 3_a^- and 3_b^-. They are for HQ60 split by the translation vector. We have not performed a similar split on all other clustering runs. Supplementary Table 1 shows for each clustering run and for each length category the following comparison to the corresponding length category for the HQ60 clustering: The first set of tables compares the different datasets considered. The second set compares various grid sizes in our clustering algorithm using the HQ60 dataset, and the third set of tables compares our clustering algorithm to the use of the Mean shift algorithm also using the HQ60 dataset. In all tables HQ60 is compared with another run which is described in the heading of the table.

For the first set of tables the first row gives the length category and the names $a, b, \ldots$ of the clusters in the run under study. The second row gives the number of rotations in the various clusters and the third row gives the percentage of these rotations that are found within clusters of the HQ60 run. Next follows a line for each cluster from HQ60 giving first the name of the cluster, the number of rotations in the cluster and the percentage of these rotations that are found in the clusters of the run under study. The remaining entries of the row correspond to the clusters

of the run under study and contain two numbers p_1/p_2. Here p_1 is the percentage of the HQ60 cluster found in the cluster corresponding to the column, and is calculated as percentage of the rotations that can actually be found within clusters from the run under study. The number p_2 is the percentage of the cluster from the run under study that is found in the cluster from the HQ60 run, and is calculated as percentage of the rotations that can actually be found within clusters from the HQ60 run.

For the second and third set of tables the third row and third column containing percentages are irrelevant (all being 100%) and are not included.

Finally, for the comparison with the Mean shift clustering the last column of the tables (with the heading *s*) collects all the minor clusters, that is, clusters that contain less than 5% of a cluster from the HQ60 run.

The overall conclusions of our clustering comparisons are as follows:

1. The following clusters, which contain over 95% of the data points, and constitute almost half of the clusters, are preserved over all the clustering runs performed: $2_a^-, 2_b^-, 3_a^- + 3_b^-$ (with Grid 71/90 and Grid 80/100 as exceptions), 3_c^- (with MS 0.3 as exception), $4_a^-, 5_a^-, 6_a^-$ (with LQ60 and MS 0.1 as exceptions), $6_b^-, 2_a^+$ (with MS 0.1 and MS 0.2 as exceptions), 6_a^+ (with LQ60 and MS 0.1 as exceptions), L_a, L_b.

2. The following clusters, which contains around 4% of the data points, are either present in a large part of the runs or they split or fuse with other clusters in the remaining cases: $2_c^-, 3_d^-, 3_e^-, 4_b^-, 5_b^-, 5_c^-, 5_d^-, 5_e^-, 4_a^+, 5_a^+, L_c, L_d, L_e, L_f$.

3. Finally, the following four clusters, which contain under 1% of the data points, recombine among themselves and with other clusters of the same length category across the different clustering runs: $3_a^+, 3_b^+, 3_c^+, 3_e^-$.

4. All the clusters of HQ60 are populated in all other clustering runs with the following unique exception: The two small clusters 2_a^+ and 6_a^+ from the dataset HQ15. This is due to the fact that HQ15 has to few data points for the algorithms to identify clusters in these two rather rare length categories.

For each of the length categories, we offer the following observations about the clusterings.

2^-: Cluster $2_a^-(16323)$ and $2_b^-(1249)$ are both very stable and are found in all other clustering runs, with the one exception, which is dataset HQ60all, where 2_a^- splits into two clusters. In fact half of the bonds from HQ60all, which is not accepted in HQ60, due to condition (33) occur in this length category and one sees that it has significant effect exactly in this length category. The small cluster $2_c^-(110)$ either re-occurs or is fused with 2_b^- or it is split into two even smaller clusters.

3^-: The clusters $3_a^- + 3_b^-(181582)$ and $3_c^-(7706)$ are very stable and are found in all other clustering runs, with the two exception, which are Grid 71/90 and Grid 80/100, where they splits into two and three clusters. The union of the two smaller cluster $3_d^-(5489)$ and $3_e^-(321)$ are found in all clustering runs (except for MS 0.03 where they are both fused with 3_c^-), but in many of the runs they fuse to one cluster and in the rest they are split into several clusters.

4^-: The biggest cluster $4_a^-(504563)$ remains in all clustering runs. Only in HQ15 is there a small cluster of size 1% split off. The smaller $4_b^-(1996)$ either remain or is split up into smaller clusters in all other runs.

5^-: The cluster $5_a^-(16499)$ remains in all runs except in Grid 80/100, where it splits off a new cluster of 25% size. The rest of the clusters $5_b^-(3672), 5_c^-(3404), 5_d^-(1893)$ and $5_e^-(295)$ are for the most parts also found in the other runs with the variation that two of them may be fused or one of them may split up into two or even three clusters.

6^-: The two clusters $6_a^-(1963)$ and $6_b^-(1312)$ remains stable over all runs with the following exceptions: In HQ60all 6_a^- splits into three clusters, one of which contains 6_b^-. This should be seen in the light of the fact, that this length category is the other one beside 2^-, where there is a (relative) significant difference of numbers of H-bonds in HQ60all and HQ60. In LQ60, MS 0.1 and MS 0.3 6_a^- splits into two, three and three clusters respectively.

2^+: Cluster $2_a^+(266)$ is found in all other clustering runs, except for HQ15. Only in the MS 0.1 and MS 0.2 runs is this cluster split into five very small clusters.

3^+: The clustering in this length category varies more when sampled over the various databases, clustering parameters and clustering method. The number of clusters vary from 2 to 6, where the interrelation between the different clusters are mixed.

4^+: The single cluster $4_a^+(6848)$ either remains or bifurcates into a number of smaller clusters.

5^+: For above half of the runs the cluster 5_a^+ splits into either two or more clusters and in the other half it remains one cluster.

6^+: The cluster $6_a^+(1325)$ is very stable over all runs with the following three clear exceptions: It is not present in HQ15 and it splits into many clusters in MS 0.1 and MS 0.2.

L: The two big clusters $L_a(242697)$ and $L_b(128332)$ are very stable and found in all runs. The same is the case with $L_d(8747)$ with the two exceptions MS 0.2 and MS 0.3, which only provides two clusters. The cluster $L_c(13726)$ either remains or it splits off a smaller cluster. The two small cluster $L_e(1221)$ and $L_f(808)$ remains stable in some runs, but may also split or fuse with some of the bigger clusters.

Supplementary Note 5. Discussion of non modal points of the long range clusters.

To understand not only the H-bonds of the mode point of a cluster, Supplementary Figure 6 shows where in rotational space the relation $\mathcal{R}_1^H = \mathcal{R}^H$ with $\mathcal{R}_1^H$ given by either of (15) or (16) holds in relation to the long-range clusters of H-bonds. The solution to each of these give rise to a 2-dimensional surface in rotational space (respectively of the form $\mathcal{R}\mathcal{R}_a^B$ and $\mathcal{R}(\mathcal{R}_a^B)^{-1}$, where $\mathcal{R}$ runs through the boundary sphere in rotational space). The extended anti-parallel beta strand, where both equations are satisfied, gives a curve on the boundary sphere in rotational space. The figure shows that cluster L_a and partly also L_d has a large overlap with the ideal anti-parallel beta strands. Cluster L_b is situated around the solution $\mathcal{R}^H = \mathrm{Id}$ to the extended parallel case. As mentioned in the Results section, clusters L_c, L_e and L_f are in between.